# English for Academic Purposes

## Approaches and Implications

**Edited by**

Paul Thompson
Giuliana Diani

# English for Academic Purposes

# English for Academic Purposes

## *Approaches and Implications*

Edited by

Paul Thompson and Giuliana Diani

Cambridge
Scholars
Publishing

English for Academic Purposes: Approaches and Implications

Edited by Paul Thompson and Giuliana Diani

This book first published 2015

Cambridge Scholars Publishing

Lady Stephenson Library, Newcastle upon Tyne, NE6 2PA, UK

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

Copyright © 2015 by Paul Thompson, Giuliana Diani and contributors

All rights for this book reserved. No part of this book may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, without the prior permission of the copyright owner.

ISBN (10): 1-4438-7439-6
ISBN (13): 978-1-4438-7439-7

# TABLE OF CONTENTS

Introduction ..... 1
PAUL THOMPSON AND GIULIANA DIANI

**PART I: CORPUS, GENRE AND DISCIPLINARY DISCOURSES**

Chapter One ..... 11
On the Phraseology of Grammatical Items in Lexico-grammatical Patterns and Science Writing
CHRISTOPHER GLEDHILL

Chapter Two ..... 43
The Role of "Lexical Paving" in Building a Text according to the Requirements of a Target Genre
GENEVIÈVE BORDET

Chapter Three ..... 79
Research Articles in Sociology: Variation within the Discipline
ŠAROLTA GODNIČ VIČIČ AND MOJCA JARC

Chapter Four ..... 103
Knowledge Construction and Knowledge Promotion in Academic Communication: The Case of Research Article Abstracts—A Corpus-based Study
MICHELE SALA

Chapter Five ..... 127
"If MSM are Frequent Testers There are More Opportunities to Test Them": Conditionals in Medical Posters—A Corpus-based Approach
STEFANIA M. MACI

Chapter Six ..... 151
Text Reflexivity in Academic Writing: A Cross-disciplinary and Cross-generic Analysis
GIULIANA DIANI

**PART II: CONTRASTIVE EAP RHETORIC**

Chapter Seven ..... 173
Interculturality in EAP Research: Proposals, Experiences, Applications and Limitations
ROSA LORÉS SANZ

**PART III: ENGLISH AS LINGUA FRANCA IN ACADEMIC SETTINGS**

Chapter Eight ..... 197
'Internationality' as a Metapragmatic Resource in Research Presentations Addressed to *English as a Lingua Franca* Audiences
LAURIE ANDERSON

Chapter Nine ..... 225
Institutional Academic English and its Phraseology: Native and Lingua Franca Perspectives
ADRIANO FERRARESI AND SILVIA BERNARDINI

Chapter Ten ..... 245
Studying ELF Institutional Web-based Communication by Universities: Comparison and Contrast with English Native Texts
GIUSEPPE PALUMBO

**PART IV: PEDAGOGICAL IMPLICATIONS IN EAP**

Chapter Eleven ..... 265
Genre, Corpus and Discourse: Enriching EAP Pedagogy
MAGGIE CHARLES

Chapter Twelve ..... 285
Text and Corpus: Mixing Paradigms in EAP Syllabus and Course Design
MARIA FREDDI

Chapter Thirteen ..... 317
Changing the Bases for Academic Word Lists
PAUL THOMPSON

Contributors ..... 343

Index ..... 349

# INTRODUCTION

PAUL THOMPSON
UNIVERSITY OF BIRMINGHAM, UK
AND GIULIANA DIANI
UNIVERSITÀ DI MODENA E REGGIO EMILIA, ITALY

Over the last two decades, there has been a prolific increase in scholarly activity in the field of English for Academic Purposes (EAP). In this growth, the notions of corpus and genre have played a central role, with important repercussions for teaching approaches. These notions derive from two approaches to the investigation of academic English: "genre analysis" and "corpus linguistics".

Genre analysis has predominantly focused on genre as text, with the aim of exploring the lexico-grammatical and discursive patterns of particular genres to identify their recognizable structural identity, or what Bhatia (1999, 22) calls "generic integrity". As Hyland observes (2012, 415), "analysing this kind of patterning has yielded useful information about the ways in which texts are constructed and the rhetorical contexts in which such patterns are used, as well as provided valuable input for genre-based teaching".

Within EAP, most genre research has used the move-analysis approach developed by Swales (1990), "which seeks to identify the recognizable stages of particular institutional genres and the constraints on typical move sequences" (Hyland 2012, 415). Substantial work has been devoted to the study of academic research genres, such as research articles, abstracts, textbooks, book reviews, book review articles, and PhD theses (see e.g. Swales 1990, 2004; Myers 1992; Bhatia 1993; Motta-Roth 1998; Bunton 2002; Lorés Sanz 2004; Kwan 2006; Diani 2012; Thompson 2012).

Genre analysis is thus largely an attempt to identify the common traits of academic language in different domains. Comparative studies have gradually shown how disciplinary domains differ from each other not only as regards specialised topics and specialist vocabulary but also in terms of lexico-grammatical characteristics and rhetorical and argumentative structures (e.g. Hyland and Bondi 2006). Even items belonging to what is

known as 'general academic vocabulary' (e.g. verbs such as *note*, *claim* or *suggest*) may be seen to vary in usage and meaning according to the specific disciplinary domain or cultural context in which they are used, thus pointing to the particular ethos of the academic community under scrutiny.

Genre analysts have been greatly assisted in these descriptions by the compilation and investigation of language corpora. The ability to analyse large quantities of data has made it possible to study the particular characteristics of different discourse domains and to investigate variation phenomena.

A significant development in these two traditions of investigating academic English is the view that genre and corpus approaches should not be considered, to borrow Charles, Pecorari and Hunston's (2009, 3) words "as opposing ideas", but rather "as constituting a continuum from top-down to bottom-up", along which researchers "situate their individual studies" (e.g. Baker 2006; Biber et al. 2007; Ädel and Reppen 2008; Charles, Pecorari and Hunston 2009; Gotti and Giannoni 2014).

The integration of genre analysis and corpus-based investigations has had a major impact on EAP pedagogy (see, for example, Weber 2001; Flowerdew 2005; Hyland 2006; Charles 2007, this volume; Cortes 2007), with important implications for teaching academic writing. This is because genre descriptions, on the one hand, support learners by providing "an explicit understanding of how target texts are structured and why they are written in the ways they are" (Hyland 2007, 151). Corpus analyses, on the other hand, encourage learners to understand academic language use, and to see the connections between language and its contexts of use. The increased familiarity of students with electronic tools for corpus analysis has contributed to the development of their language awareness (e.g. Bondi 1999) and promoted learner autonomy (e.g. Lynch 2001).

## Contents of the volume

Many of the issues outlined above are investigated in the chapters of this volume. The contributions are arranged into four parts, which highlight how corpus linguistics and genre analysis can work as complementary approaches. Pedagogical implications are also discussed in some detail as the research described here not only aims to investigate features of EAP but also to translate them into classroom applications. The first part presents corpus-based research into EAP at the lexicogrammatical and genre levels, with papers whose focuses range from issues related to patterning and phraseology through to papers focusing on

the language practices of specific disciplines and research genres. The second part is devoted to intercultural EAP research. The third part includes research on English as a Lingua Franca in academic communication. Finally, the last part addresses the relationships between corpus, genre and pedagogy in EAP, with an emphasis on implications and applications.

## Overview of the chapters

The first two chapters of Part I "**Corpus, Genre and Disciplinary Discourses**", focus on lexico-grammatical patterning in written academic discourse. The opening article, by **Christopher Gledhill**, investigates how collocation and phraseology, from a lexicogrammar perspective, are relevant to EAP. He sets out a method of textual analysis which exploits the phraseological behaviour of grammatical signs. The chapter provides evidence that grammatical items can be shown to be stable elements in relatively predictable but also productive cascades of expression. His findings suggest that the identification of such extended lexicogrammatical patterns are a key feature in the systematic analysis of EAP texts.

The second chapter, by **Geneviève Bordet**, also treats the phenomenon of collocation as fundamental to the study of language, and genre analysis in particular. Her analysis provides evidence that each genre highlights specific disciplinary strategies based on the use of textual collocational variations, or "collocational chains". Her results show that these chains contribute to the perception of the text as both coherent and persuasive.

The second trend of Part I is represented by four chapters centring on language variation across disciplines in different academic genres: research articles, book review articles, abstracts, conference posters. **Šarolta Godnič Vičič** and **Mojca Jarc** explore language variation in the genre of research articles within sociology. Three functional key words, *among*, *between* and *these*, were selected for analysis as the most salient in all the corpora analysed. Their study demonstrates that there is intradisciplinary variation in the sociological research articles that is not due to the stylistic flexibility with which authors use language. They suggest that the differences between the preferred meanings and values found in the corpora may be attributed to differences in the research focus and theoretical positionings of the authors, the methodologies they use, and also the niches occupied by the journals.

**Michele Sala**'s chapter examines the way knowledge is conceptualized and grammatically constructed in research article abstracts covering four

disciplinary areas (linguistics, law, economics and medicine). The analysis focuses on linguistic items which are used to portray the construction of disciplinary knowledge, to introduce concepts and methods, and to represent evidence and its interpretation. His study shows a consistent use of *research*, *cognition* and *discourse acts* in the disciplinary corpora analysed.

**Stefania M. Maci** investigates how conditional constructions are employed in the discourse of medical posters. Her results show that, in this genre, the uses of conditionals reflect the reasoning process of the hard sciences: they can either convey 'facticity' or 'refocusing'. Through the analysis, her study reveals that in the case of 'facticity', facts and results are reported according to the conditions they are associated with and are expressed in the 'Method' and 'Results' sections. As regards factual *if* clauses, their pragmatic role is indicated by the information ordering structuring: prediction seems to be realised with the fronting of protasis, whereas the *if and only if* condition appears to be constructed by means of delaying.

In the next chapter, **Giuliana Diani** explores reflexive phraseology across academic genres and disciplines. Employing a corpus-based approach, the study focuses on how metatextual phraseological units vary across academic research articles and book review articles and academic disciplines (business and economics). Her study shows that phraseological units can be very helpful signals for the analysis of generic and argumentative structure of academic writing. Her findings also demonstrate that convergences and divergences between closely related disciplines and genres help to differentiate different forms of disciplinary discourse.

With Part II of the volume, **"Contrastive EAP rhetoric"**, our attention turns to an investigation of intercultural EAP research. **Rosa Lorés Sanz** illustrates a methodological approach for intercultural research in the use of written academic genres in English by non-native (Spanish) academics, which involves corpus analysis, genre analysis and intercultural rhetoric as central methodological frameworks. She presents some of the findings that have resulted from its application to the exploration of non-native (Spanish) use of EAP. Special emphasis is also made on the advantages and the limitations faced by this methodological approach. The application of a cross-cultural approach to the design of teaching materials and the implementation of EAP courses is also discussed.

Part III "**English as lingua franca in academic settings**" is devoted to research on English as Lingua Franca (ELF) in academic communication. **Laurie Anderson** investigates the pragmatics of academic ELF

communication by examining the role that the thematization of self and other identities in terms pertinent to membership in an international community of scholars plays in peer-to-peer interaction among academics from different national backgrounds. Her study shows that in peer-to-peer interaction with colleagues from different national and linguistic backgrounds, scholars exhibit a particular understanding of international academia, reflecting both the geographical/geopolitical and institutional characteristics of the setting in which the presentations were made. The chapter concludes with a discussion on the extent to which the thematization of 'internationality' is rooted in the specific aims of the genre analysed and the extent to which it is instead a more pervasive aspect of academic communication in ELF settings.

In the next chapter, **Adriano Ferraresi** and **Silvia Bernardini** report on an ongoing project focusing on institutional (vs. disciplinary) academic English as a Lingua Franca as it is used in university websites worldwide to present degree programme descriptions and syllabi and to provide information on a wide range of administrative and organisational matters. They investigate the use of phraseology in native and ELF varieties represented in a 90-million word corpus of institutional academic texts in English. Their findings reveal that ELF university homepages display (phraseological) patterns which are only partially consistent with previous studies of non-native language, and that these deviations might or might not derive from conscious strategies to target an international audience.

In the final chapter of this section, **Giuseppe Palumbo** investigates features of comparable sets of texts written in ELF by universities and in two national varieties of English, with a view to identifying the way the texts construct their respective profiles at the morpho-syntactic level and realize their main, shared function. His study shows that the non-native, ELF texts present similarities and differences from comparable native texts. Some differences between the ELF set and the two native sets concern the use of verbs, while similarities regard the purely structural aspects (such as the generalised tendency to use premodification in noun phrases) and the use of patterns pointing to the adoption of similar signals of stance or engagement, such as the heavy personalization of the discourse through the use of pronouns ("we"/"you" as opposed to "the university"/"students"). His analysis highlights a certain homogeneity between the non-native and native sets with regard to their structural make-up.

Part IV **"Pedagogical implications in EAP"** turns attention to EAP pedagogy. **Maggie Charles** discusses how corpora can be used to enrich EAP pedagogy by facilitating the study of genre and discourse features in

academic writing. She illustrates this through two approaches. The first uses traditional paper-based materials derived from prior analysis of a corpus and shows how pedagogical tasks can contribute to raising student/learner awareness of the variability of genres. The second approach uses a hands-on method, in which the students/learners are responsible for building and investigating their own corpora and shows how they can make use of their corpora to examine discourse functions in their own discipline.

**Maria Freddi**'s contribution focuses on EAP reading pedagogy. The chapter reports on the research informing the development of a course taught by the author at her home university, aimed at undergraduate Humanities students and entitled *Reading Skills in English for the Humanities*. It explores ways in which insights from corpus approaches to academic English combined with genre theory can be brought to bear on the design of the course syllabus and argues for a pedagogically targeted mix of the various paradigms under consideration.

In the last chapter of the volume, **Paul Thompson** explores the lexis of academic lectures through analyses of frequency and range of items within a corpus. He tests the coverage of the Academic Word List (Coxhead 2000) and General Service List against three other options and concludes that Paul Nation's 2K word list, based on BNC frequencies, provides a better indication of the most frequent items in academic lectures than does the General Service List. He then develops a specialised academic lecture listening word list made up of 200 items. Finally, he presents some corpus exploration activities, based around the new word list, which access the British Academic Spoken English (BASE) corpus in the open Sketch Engine interface, that can be used with learners preparing for the challenges of listening to lectures in English.

The various analyses collected in this volume provide a rich overview of the methods of investigation of EAP, the tools and the approaches, bringing together, to differing degrees, two complementary strands of linguistic investigation – corpus analysis and genre analysis. They demonstrate how the wealth of data made available through corpus compilation and searchable through query tools have enabled scholars to identify and give clear descriptions and examples of central concepts in EAP research.

## References

Ädel, Annelie, and Randi Reppen. 2008. *Corpora and discourse: The challenges of different settings*. Amsterdam: John Benjamins.

Baker, Paul. 2006. *Using corpora in discourse analysis*. London: Continuum.

Bhatia, Vijay K. 1993. *Analysing genre. Language use in professional settings.* London: Longman.

—. 1999. Integrating products, processes, processes and participants inprofessional writing. In *Writing: Texts, processes and practices*, ed. Christopher N. Candlin and Ken Hyland, 21-39. London: Longman.

Biber, Douglas, Ulla Connor, and Thomas Upton. 2007. *Discourse on the move*. Amsterdam: John Benjamins.

Bondi, Marina. 1999. Language awareness and EFL teacher education. In *English teacher education in Europe: New trends and developments*, ed. Pamela Faber, Wolf Gewehr, Manuel Jiménez Raya and Antony J. Peck, 91-107. Frankfurt: Peter Lang.

Bunton, David. 2002. Generic moves in PhD thesis introductions. In *Academic discourse*, ed. John Flowerdew, 57-75. London: Longman.

Charles, Maggie. 2007. Reconciling top-down and bottom-up approaches to graduate writing: Using a corpus to teach rhetorical functions. *Journal of English for Academic Purposes* 6(4): 289-302.

—. this volume. Genre, corpus and discourse: Enrich EAP pedagogy.

Charles, Maggie, Diane Pecorari, and Susan Hunston. 2009. *Academic writing: At the interface of corpus and discourse*. London: Continuum.

Cortes, Viviana. 2007. Genre and corpora in the English for academic writing class. *ORTESOL Journal* 25: 9-16

Coxhead, Averil. 2000. A new academic word list. *TESOL Quarterly* 34(2): 213-238.

Diani, Giuliana 2012. *Reviewing academic research in the disciplines: Insights into the book review article in English*. Rome: Officina Edizioni.

Flowerdew, Lynne. 2005. An integration of corpus-based and genre-based approaches to text analysis in EAP/ESP: Counting criticisms against corpus-based methodologies. *English for Specific Purposes* 24(3): 321-332.

Gotti, Maurizio, and Davide S. Giannoni. 2014. *Corpus analysis for descriptive and pedagogical purposes: ESP perspectives*. Bern: Peter Lang.

Hyland, Ken. 2006. *English for academic purposes: An advanced resource book*. London: Routledge.

—. 2007. Genre pedagogy: Language, literacy and L2 writing instruction. *Journal of Second Language Writing* 16: 148–164.

—. 2012. English for academic purposes and discourse analysis. In *The Routledge handbook of discourse analysis*, ed. James Paul Gee and Michael Handford, 412-423. London: Routledge.

Hyland, Ken, and Marina Bondi. 2006. *Academic discourse across disciplines*. Bern: Peter Lang.

Kwan, Becky S.C. 2006. The schematic structure of literature reviews in doctoral theses of applied linguistics. *English for Specific Purposes* 25(1): 30-55.

Lorés Sanz, Rosa. 2004. On RA abstracts: From rhetorical structure to thematic organisation. *English for Specific Purposes* 23(3): 280-302.

Lynch, Tony. 2001. Promoting EAP learner autonomy in a second language university context. In *Research perspectives on English for academic purposes*, ed. John Flowerdew and Mathew Peacock, 390-403. Cambridge: Cambridge University Press.

Motta-Roth, Désirée 1998. Discourse analysis and academic book reviews: A study of text and disciplinary cultures. In *Genre studies in English for academic purposes*, ed. Inmaculada Fortanet, Santiago Posteguillo, Juan C. Palmer and Juan F. Coll, 29-58. Castelló de la Plana: Universitat Jaume I.

Myers, Gregory. 1992. Textbooks and the sociology of scientific knowledge. *English for Specific Purposes* 11(1): 3-17.

Swales, John. 1990. *Genre analysis. English in academic and research settings*. Cambridge: Cambridge University Press.

—. 2004. *Research genres. Explorations and applications*. Cambridge: Cambridge University Press.

Thompson, Paul. 2012. Thesis and dissertation writing. In *Blackwell handbook of English for specific purposes*, ed. Brian Paltridge and Sue Startfield, 283-300. Oxford: Wiley-Blackwell.

Weber, Jean-Jacques. 2001. A concordance- and genre-informed approach to ESP essay writing. *English Language Teaching Journal* 55(1): 14-20.

# PART I

## CORPUS, GENRE AND DISCIPLINARY DISCOURSES

# CHAPTER ONE

# ON THE PHRASEOLOGY OF GRAMMATICAL ITEMS IN LEXICO-GRAMMATICAL PATTERNS AND SCIENCE WRITING

CHRISTOPHER GLEDHILL
UNIVERSITÉ PARIS DIDEROT, FRANCE

## 1. Introduction

In this chapter I examine the role of grammatical items in lexico-grammatical patterns (or 'LG patterns', for short). In previous work I examined the collocational patterns of individual grammatical items in a particular genre (the cancer research article, Gledhill 1995, 2000a, 2000b). In these studies, I demonstrated that individual functional words (such as 'and' in Titles, 'but' in Abstracts, 'to' in Introductions and so on) have a non-random distribution in these texts, since these words are 'statistically salient' or 'key' in these different parts of the research article. I then went on to examine in detail the phraseological behaviour of these items in each subsection, arguing that, contrary to what one might think, each grammatical item enters into a very restricted, predictable set of phraseological patterns, according to the type of text being analysed. It is often thought that grammatical items do not enter into collocational relations, since they can 'be used anywhere' and thus can 'collocate with anything'. However, one of the findings of my work has been to argue that grammatical items have a highly restricted phraseology in specialised discourse, a feature which makes them ideal targets for analysis, since an analysis of the distribution and behaviour of function words can in effect be seen as a preliminary analysis of the fundamental stylistic and phraseological properties of a particular text type.

In this chapter, I use similar methods and I make similar claims. However, my focus here is somewhat different. In this study, I concentrate on longer stretches of wording, and in particular I am interested here in examining discontinuous stretches of text or what Renouf and Sinclair (1991) have called "collocational frameworks". A discontinuous stretch of text is a short sequence of words such as *a(n) * *-ed in* in which only a few selected kinds of lexical item (marked by *) can fit meaningfully into the lexical pattern (in this case '*A substance found in... A cytokine implicated in'...* etc.). What I find interesting about such sequences is that they often correspond to short extracts of very highly specialised discourse. The main hypothesis I wish to test here is that by searching for a given sequence of grammatical signs, it is possible to identify a regular pattern of discourse, provided that this pattern is looked for in a relatively coherent body of texts (i.e. an electronic archive or corpus). Furthermore, I would suggest that by looking at discontinuous sequences in this way, the corpus analyst should be able to identify some of the most characteristic features and functions of that genre in an efficient and systematic manner. In this chapter, I look at examples taken from a corpus of research articles and their corresponding abstracts (referred to below generally as RAs), as well as 'journalistic accounts' of the same research (referred to as JA). However, as I point out in the data analysis below, although many discontinuous stretches are highly regular and recurrent in the particular texts I am interested in here, some kinds of writing (notably scientific journalism) also involve 'hybrid' patterns, i.e. an original blending of two or more patterns that are commonly found in other types of discourse.

In this contribution I use the term "grammatical item" (also known as closed-class item, function word, small word, stop-word, etc.) in contrast to "lexical item", and I assume that the difference between the two is that grammatical items belong to a relatively closed class of high-frequency, polyvalent words with relatively abstract meanings, such as *auxiliaries, conjunctions, determiners, grammatical adverbs, prepositions* and *pronouns*. Until recently, grammatical items as an entire class have not received much attention in English for Specific Purposes and corpus-based genre analysis. Indeed, it has often been supposed that grammatical items are of little interest in text linguistics because they can "go with anything" (i.e. collocate with any lexical item, appear in a vast range of grammatical structures, occur in a uniform way across almost all text types, etc.). Thus in the early days of lexicometrics and Natural Language Processing (NLP), specific procedures were developed to filter these words out from the analysis (for example, Smadja 1993 introduced a well-known method of extracting collocations from the Hansard bilingual corpus, but only after

a process of automatically filtering out the function words). To a certain extent the concept of the stop-word is still widespread, and entirely understandable: analysts are interested in getting at what they consider to be important data (equating this with lexical items, content-words, terminology), while the large quantities of apparently ubiquitous grammatical items which fill most types of text appear to be redundant 'noise'.

However, since the advent of corpus linguistics (and particularly the use of corpora by grammarians), there has been a growing body of evidence to suggest that grammatical items enter into collocational relations that are just as interesting and revealing as lexical items, and thus have an important role to play in the phraseological patterning of texts. Renouf and Sinclair (1991) and Renouf (1992) most notably argued that grammatical items are the building blocks of idiomatic language, and coined the term "collocational framework" to refer to such productive sequences as *a(n) X of (a [dash, handful, smattering] of)*. More recently, and in a way that mirrors the current move to rehabilitate 'junk DNA',[1] some researchers in NLP now recognise the importance of functional words in automatic text recognition, terminology extraction and other corpus-based applications (Meyer 1988; Riloff 1995; Vergne 2004). Similarly, collocational frameworks and related notions such as "bundles", "clusters", "n-grams" and so on, have become an accepted part of the descriptive apparatus of corpus-based applied linguistics, and many researchers have examined the distribution and collocational behaviour of specific grammatical items as they occur in specialised corpora (Luzón Marco 1999, 2000; van der Wouden 2001, 2007; Biber et al. 2004; Cheng et al. 2008; Hyland 2008; Bordet 2011) as well as the role of grammatical sequences and frameworks in discourse analysis, language learning and evaluation (Hasselgren 2002; Groom 2005, 2010; Scott and Tribble 2006; Biber and Barbieri 2007; Lee et al. 2008).

But although the study of function words has now become an accepted part of the corpus-based approach, the notion that grammatical items can be the focus of phraseological analysis still requires some theorisation within a broader analytical framework. In addition, it seems to me that for most observers, the idea that grammatical items can be the starting point for textual analysis is still not obvious. Therefore, before embarking on an analysis of discontinuous lexico-grammatical patterns, I set out in the first half of this chapter some of the arguments for studying grammatical items from the point of view of corpus-based genre analysis.

## 2. Why study grammatical items?

### 2.1. Grammatical items have collocations

Not all analysts accept that grammatical items enter into collocational relations. For example, in one well-known British dictionary of linguistics, "collocation" is defined as the co-occurrence of lexical items, while grammatical items are explicitly stated as having no collocational relations:

> collocation, n. A term used in lexicology by some (especially Firthian linguists) to refer to the habitual co-occurrence of individual lexical items […] Some words have no specific collocational restrictions - grammatical words such as *the, of, after, in* [...] Another important feature of collocations is that they are formal (not semantic) statements of co-occurrence [...] (Crystal 2008, 86-87)

This definition seems to represent a commonly-held view among linguists. However, I feel that Crystal rather misrepresents the way Firth (1957) would have understood the term, or at least as Firth's successors understand it. From a Firthian point of view, every single sign in the language (whether lexeme or morpheme) has a consistent and contrastive context of use. In other words, each sign is used in a consistent and contrastive set of linguistic co-texts (e.g. the preposition *of* typically has as a complement the noun *course*) and each sign is used in a consistent and contrastive set of situational contexts (e.g. *of course* tends to be used as an informal, concessive adjunct). The term "context" is clearly central to this approach, and in typical Firthian fashion it is used to indicate three things at the same time: a) the "the co-occurrence of forms within the same stretch of text" (usually what is meant by "co-text"), b) the immediate "context of situation", and c) the broader "context of culture". The point about co-text and context is that they essentially shape the meaning of the linguistic sign, since signs are mutually dependent on their typical collocational partners in discourse, as Firth puts it:

> The collocation of a word or a 'piece' is not to be regarded as mere juxtaposition, it is an order of mutual expectancy. The words are mutually expectant and mutually prehended. (Firth 1957 [1951], 181)

This kind of definition sets itself against an "essentialist" or "semantic trait" approach to meaning. Thus, it is claimed that the meaning of a word such as *of* can only be seen as a composite of its particular uses, which

depend partly on its co-occurrence with *course* in *of course* and partly on other uses, such as its co-occurrence with a nominal post-modifier referring to people in terms of subjective, usually high-mindedly positive qualities (*a man / woman of [action, honour, humility*, *steel, quality]*) and so on. Sinclair (1991) argued that these and the other typical lexico-grammatical patterns associated with *of* contribute to our general understanding of this rather idiosyncratic preposition. As he puts it:

> Most everyday words do not have an independent meaning, or meanings, but are components of a rich repertoire of multi-word patterns that make up text (Sinclair 1991, 108).

## 2.2. Grammatical items have different distributions in different text types

Many linguists would agree that different varieties of language exploit different lexical and grammatical resources, an assumption which lies behind the multi-factorial method of register analysis developed by Biber et al. (2002). Yet it is surprising to see how few studies of a particular text type begin by setting out the relative distribution of grammatical forms, especially grammatical items. In this chapter, I examine the role of grammatical items in their local contexts (as the most recurrent, pivotal items in lexico-grammatical patterns). But before examining their local role, it is important to realise that grammatical items (indeed all items, whether functional or lexical) have a particular distribution in different texts, and even within different subsections of texts. The reason for this is that that, as shall be shown in the following sections, if the occurrence of a particular grammatical item is statistically more 'salient' or 'key' in a particular text type or subsection of a text, this is because this item is pivotal in those lexical patterns which have an important role or discourse function to play in that text. One example of this will suffice here: in Gledhill (2000) I showed that the word 'to' is an statistically significant item in cancer research article introductions. One reason for this, it would seem, is that 'to' is a pivotal element in post-posed attributive clauses such as *'it is (important, necessary, possible) to (assess the cell differentiation at this stage, construct a series of structures, identify TAAs, repeat measurements)...*' but also in passive non-finite projecting clauses such as '*(HPV 16 E6, hyperphasis, metabolic inc-cells, is (known, likely, thought...) to (be involved, be a major factor, determine celle cycle) in' etc*.' (from the Pharmaceutical Sciences Corpus, Gledhill 2000, 151). Examples such as these demonstrate three principles: 1) grammatical items collocate with other lexical and grammatical items (note the discontinuous

sequence *it is X to* in the first pattern and *X is Y-ed to Z in* in the second pattern), 2) each lexico-grammatical pattern expresses a specific discourse function (in the first case, 'strongly stating the case for a clinical methodology' and in the second case 'tentatively proposing a biochemical explanation'), and 3) the most systematic way of identifying these patterns is, in my view, to compare the relative distribution of grammatical items across different text corpora.

As mentioned in the introduction, in Gledhill (1995, 2000a, 2000b) I attempted to show that grammatical items have a particular distribution across the different rhetorical sections (Titles, Abstracts, Introductions, etc.) of 150 research articles (RAs) in the field of cancer research (the Pharmaceutical Sciences Corpus, PSC). At the time, I did not have access to a specialised reference corpus in English but later I used the British National Corpus (BNC) as a reference corpus. For example, the following tables (1, 2 and 3) present the results of a "keyword" comparison using AntConc (Antony 2002). In order to obtain these results, AntConc first creates a word list for each corpus (the BNC has 100,520,565 tokens with 448,005 types, and the PSC has 1896869 tokens with 48,537 types).[2] AntConc then calculates a score for each word by comparing a particular item's percentage chance of occurring in the study corpus as opposed to its chances of occurring in the reference corpus. For example, in the PSC *were* occurs 17,968 times divided by 1,896,869 tokens (=0,0961 or roughly 0.96%), whereas in the BNC *were* occurs 306,801 times divided by 100,520,565 tokens (=0,00305 or roughly 0.30%). Such a large percentage difference shows the extent to which certain function words, such as *were,* can have consistently different distributions across text types. It is of course important to be able to judge what is meant by "large percentage difference", and the "Keywords" module of AntConc is an important stage in this analysis. Keywords compares the relative percentages of items as they occur in the study corpus and reference corpus, and then assigns a rank to each item according to a statistical test (for details see Scott and Tribble 2006).[3]

The tables set out the first 20 keywords in the PSC (ranked by decreasing keyword score) as compared with the BNC (Table 1) and the first 20 key grammatical items in the PSC as compared with the BNC (Table 2).

| Rank | Item | Freq. in PSC | Keyword score(vs. BNC) | Rank | Item | Freq. in PSC | Keyword score (vs. BNC) |
|---|---|---|---|---|---|---|---|
| 1 | & | 6071 | 48432.645 | 11 | Table | 2231 | 8976.632 |
| 2 | patients | 8563 | 36825.897 | 12 | clinical | 1812 | 8445.988 |
| 3 | et | 4540 | 22956.671 | 13 | cell | 2165 | 8335.861 |
| 4 | al | 4544 | 21851.859 | 14 | min | 1447 | 7897.745 |
| 5 | study | 5791 | 19067.584 | 15 | Fig | 2126 | 7643.490 |
| 6 | cells | 3911 | 16769.641 | 16 | cases | 3053 | 7609.082 |
| 7 | were | 17968 | 15873.864 | 17 | patient | 2249 | 7373.035 |
| 8 | results | 3664 | 11613.943 | 18 | studies | 2467 | 7307.815 |
| 9 | treatment | 3212 | 10486.346 | 19 | significant | 2410 | 6703.943 |
| 10 | of | 82182 | 9404.549 | 20 | tissue | 1373 | 6402.795 |

**Table 1. First twenty keywords in the Pharmaceutical Sciences Corpus (PSC) vs. the BNC.**

| Rank | Item | Freq. in PSC | Keyword score(vs. BNC) | Rank | Item | Freq. in PSC | Keyword score (vs. BNC) |
|---|---|---|---|---|---|---|---|
| 7 | were | 17968 | 15874.251 | 228 | these | 3859 | 1652.307 |
| 10 | of | 82182 | 9404.238 | 236 | However | 1702 | 1627.852 |
| 38 | with | 21063 | 5318.315 | 247 | due | 1203 | 1595.017 |
| 45 | in | 47809 | 4915.890 | 260 | may | 4165 | 1529.965 |
| 58 | and | 61723 | 3824.316 | 261 | The | 16217 | 1528.736 |
| 91 | during | 2581 | 2971.304 | 347 | Therefore | 461 | 1250.765 |
| 117 | between | 4148 | 2521.590 | 354 | or | 9940 | 1230.950 |
| 124 | In | 6005 | 2464.055 | 361 | after | 3479 | 1213.382 |
| 163 | vs | 396 | 2114.845 | 364 | was | 20876 | 1208.869 |
| 189 | versus | 456 | 1888.239 | 519 | both | 2313 | 921.387 |

**Table 2. First twenty grammatical keywords in the Pharmaceutical Sciences Corpus (PSC) vs. the BNC.**

Analysts familiar with keyword lists will have little difficulty in interpreting these data. The keywords in Table 1 show some of the major textual features of the PSC (such as the presentation of *data* in *Tables* and *Figures*) as well as the topical preoccupations of the PSC (the nominal expression of material processes, e.g. the *study* of *cells, cases* or *groups* of different *ages* and at different *times*, the *treatment* of *patients*, the *use* of drugs / *treatments* and the verbal expression of communicative or perceptive processes *significant results found* or *reported*). Similar

comments can be made about the key grammatical items in the PSC (Table 2): they are predominantly prepositions and coordinators (*of, and, or,* typically involved in elaborate nominal post-modifiers in the PSC) or prepositions involved in adjuncts / post-modifiers expressing cause (*due to*), accompaniment or manner (*with*), temporal extent (*after, between, during, in*) and comparison (*between, versus, vs.*). Table 2 also shows the typical markers of cohesion which we might expect to find in elaborate written discourse, such as pronouns / determiners (*both, The, these*) and sentence-initial conjuncts (*However*, *Therefore*). Other items are perhaps less obviously typical of written discourse, but Table 2 shows that they are salient in science writing: *may* (the preferred modal verb for "hedging", especially in Discussion sections of the PSC) and *were* (usually an auxiliary expressing the past passive in Methods sections).

As mentioned above, my initial description of the PSC was an intra-varietal analysis, conducted in order to establish the main differences between the different rhetorical sections of the research article and the PSC corpus as a whole (Titles, Abstracts, Introductions, Methods, Results, Discussions). I shall not repeat these data here, but for illustrative purposes, the following Table 3 sets out the main results for the first ten key grammatical items across each sub-section of the PSC:

| Rank | Rhetorical Section of the PSC | | | | | |
|---|---|---|---|---|---|---|
| | Title | Abstract | Introduction | Methods | Results | Discussion |
| 1 | of | but | been | were | no | that |
| 2 | for | these | has | was | in | be |
| 3 | on | of | have | at | did | may |
| 4 | and | there | is | then | not | is |
| 5 | in | in | such | for | had | our |
| 6 | - | was | can | each | after | in |
| 7 | - | that | it | and | there | not |
| 8 | - | did | we | from | the | this |
| 9 | - | who | of | after | when | we |
| 10 | - | both | to | with | all | have |

**Table 3. First ten grammatical keywords in the six main sub-sections of the PSC.**

Although general comparisons (Tables 1 and 2) give a good idea of the general features of the PSC, Table 3 shows the extent to which there is also much variation within the research article genre itself. In fact there is so much internal variation that some items which are statistically salient in their respective sub-sections of the PSC, are also more typical of the

general language (BNC) when compared with the PSC as a whole (in particular: the pronouns *we, our* in Introductions and Discussions, the modal *can* which is only salient in Introductions, the item *to* which is also salient in Introductions, as mentioned above). In Table 3, I have indicated the items which stand out in relation to the other sub-sections of the RA (by underlining). I shall not go into a detailed analysis of these data here. It is sufficient to note that over half the keywords in the Introductions (*been, has, such, can, it, to*) and Methods (*were, at, then, each, from, with*) are only specifically "key" in these sections, a result which suggests that these sections have a specific phraseology which is quite unlike the rest of the research article (although these items are of course not exclusive to these sections).

## 2.3. Grammatical items are pivotal elements in lexico-grammatical patterns

In the previous section, I showed that grammatical items do not have an even distribution across text types, and that their distribution varies considerably, even within the same text type. In this section I argue that the identification of "key" grammatical items can be seen as a useful first stage in the search for longer stretches of phraseology. Some authors (notably Hunston and Francis 2000) use the term "lexical pattern" to refer to regular multi-word units which do not necessarily correspond to the traditional constituents of the clause. In this chapter, (and elsewhere, Gledhill 2011) I refer to such sequences as "lexico-grammatical" (LG) patterns, in an attempt to make it clear that in any multi-word phrase at least one grammatical item (or grammatical structure) is a permanent or "pivotal" element around which the rest of the phrase is built.

In order to illustrate this notion, let us return to the particular case of Abstracts in cancer research articles (in the PSC there are 400 Abstracts = 123,296 words or 6.5% of the corpus). As mentioned above, the first ten grammatical keywords in this sub-section are (in order of rank) *but, these, of, there, in, was, that, did, who, both.* Out of context, it is not clear what patterns of usage these items might represent. An item such as *that* can be used in many different lexico-grammatical contexts (conjunction, pronoun, determiner etc.), and it is therefore necessary to analyse each item separately, not only within Abstracts, but also contrastively, in the rest of the research article. This is not an easy task, not least because the analysis of grammatical items usually generates a vast amount of data. However, I would suggest that the task is simpler when looking at a specialised genre than for the general language. For example, in the 1st edition of the Collins

Cobuild dictionary (Sinclair 1995), there are 19 entries for *of* (not including idiomatic uses), whereas in the PSC (Gledhill 2000,142-149) the number of patterns varies between 3 (Titles, Abstracts) and 5 (Introductions). Even so, it is still difficult to represent this kind of data, and I will not repeat the analysis of each of these patterns here, largely because this type of detailed analysis requires long lists of concordances. However, in Gledhill (1995) I suggested one way of resolving the problem of data representation, which I called the "collocational cascade". An example of this is set out in the following figure:

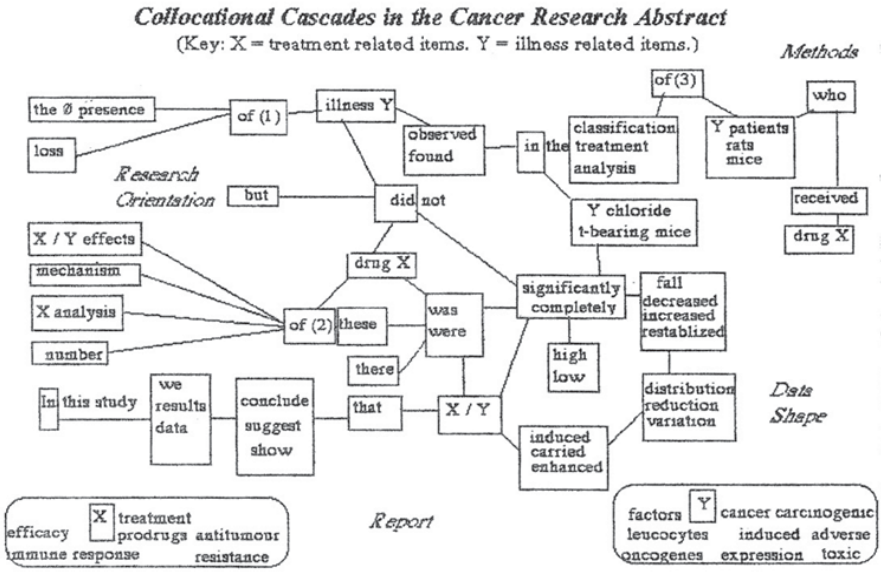



**Figure 1. Collocational cascades in Cancer Research Abstracts.**[4]

I would argue that "collocational cascades" are an efficient way of summarising the most salient phraseology of a particular text type. Thus in Figure 1 we can see that the outstanding or salient phraseology of Abstracts involves a specification of the general shape of data (as evidenced by lexico-grammatical patterns involving *in, of, that, there*), statements about who or what was affected by various treatments (the LG patterns around *in*, *of, who*) and the extent to which an effect was or was not observed (the LG patterns involving *but, did, not, was, were*). Unlike diagrams representing collocational networks (Williams 1998),

collocational cascades do not represent a formal or statistical relationship between lexical items. Rather, the cascade is broadly meant to be read from left to right, as an informal representation of interlocking lexico-grammatical patterns which, as the cascade metaphor suggests, fall or lead on from one choice of expression into another further on in the clause. In Figure 1, each grammatical item in the cascade (*but, did + not, in, of, that, there, these, who,* but not *both*) is linked to one or more of the main patterns as observed in the Abstracts sub-corpus (not counting of course the many sub-patterns or variants of these patterns). We can also see in the diagram that some items appear more than once. Thus, the diagram shows that there are 3 (main) patterns involving the preposition *of* in Abstracts: (1) a quantification (*loss / presence*) of a (usually post-modified) disease-related item (*cancer cells, carcinogenic factors, leucocytes...*), (2) an observation, quantity or facet (*amount*, *analysis*, *effect*) of a treatment-related item *(antitumour response, immune response, prodrug*), and (3) an extended pattern involving a reduced relative clause (expressing a mental / empirical-oriented process: *found*, *observed*) qualified by a complex nominal (expressing a research-oriented process: *in the + analysis / classification / treatment of + disease-related item*). While these patterns are clearly prevalent in Abstracts, they are also clearly typical of the complex (post-modifying) nominals analysts have come to expect in academic and scientific writing in general. In addition, in the following section, we see that pattern 4 has a slightly different realisation in journalism (examples 4g, 4h and 4i).

One of the defining features of collocational cascades is that the patterns they represent are all related (each grammatical and lexical item is linked indirectly to one or more other items, creating a complex, although sometimes incomplete chain of patterns). I would claim that the grammatical items in the cascade are "pivotal", that is to say they are used consistently in each of these patterns. The lexical items on the other hand represent a "paradigm", in that they usually represent semantic abstractions or families of related lexical items rather than specific examples (as in *of* pattern 1: *loss / presence of* + Y, where Y is the name of a specific disease-related item such as *leucocytes*). In addition, collocational cascades have a certain directionality. What I mean by this is that the cascade as a whole represents the general way in which information is structured within Abstracts. For example, expressions and phrases which are research- or report-oriented (*In this study, we conclude that...*) as well as empirical observations (*loss / presence of item Y*) appear in theme (clause-initial) position in this diagram, whereas clinical methods

(*item Y who received item X*) and results (*did not significantly fall / increase...*) appear in rheme / clause-final position.

I make no claim here about the linguistic status of collocational cascades; I primarily see them as a way of depicting the outstanding phraseology of a particular text type. However, I would suggest that this kind of representation does capture something of the social or psychological reality of this kind of formulaic language, in which members of the discourse community recognise that they write in "chunks" and "formulae", and claim to "skim" research articles before deciding whether to pore through them line-by-line. These notions, as well as non-linear processing, predictive text analysis and lexical priming, have become important themes in applied linguistics (de Cock 1998; Simpson 2004; Hoey 2005). However, I will not dwell on these issues here. The more general point I am making is that lexico-grammatical patterns are significantly recognisable within a particular text type, and that LG patterns can be seen as parts of a broader set of interlocking collocational cascades within that particular type of discourse.

## 3. Extended lexico-grammatical patterns

So far, I have argued the case for seeing grammatical items as key elements in the corpus-based analysis of genres, since these words are the pivotal building blocks of lexico-grammatical patterns. In this section, I attempt to establish whether longer stretches of LG patterns can be identified on the basis of corpus analysis, and my particular focus here is on patterns involving extended (and usually discontinuous) sequences of grammatical items. Previous work on collocational frameworks has usually focused on the collocation of two grammatical items within a pre-defined window of words (Renouf and Sinclair 1991; Cheng et al. 2008; Groom 2010). Here, on the other hand, I am interested in identifying patterns in which at least one grammatical sign is bound between two other grammatical items, such as *the* * * *of* **s* (where * is a wild-card representing one lexical item of any length, * * are contiguous lexical items, and [* *of*] or [**s*] stand for lexical items with an intervening grammatical item or an attached grammatical morpheme). My hypothesis is that discontinuous sequences of grammatical items are particular to specific genres, and that when they can be observed with sufficient regularity, they provide good evidence for the existence of the typical lexico-grammatical patterns of that text type. In other words, given two sequences, such as a) and b) below (both sequences are complete sentences), it should be possible to predict whether they belong more or

less to the typical discourse of a research article (RA) or a journalistic article (JA), and within these sequences it should be possible to identify those patterns that are typical of the genre and those which are 'merely' local innovations:

a) * * * is a * *of * *s and some *s * that *ly * of *s are *ed.
b) *s * a * *er to *ing an * * * and * * of * of the most * and * *s.

In order to test the 'extended pattern' hypothesis, I have examined a sample of complete sentences taken from research articles on cancer cachexia (two of which authored or co-authored by Professor Michael Tisdale, Aston University, and both included in the Pharmaceutical Sciences Corpus, PSC) and from a selection of journalistic articles which all refer to this research as a "breakthrough".[5] In the following sections I look at the initial sentences from two research articles written by Michael Tisdale and his colleagues (RA1: *Trends in Pharmaceutical Sciences* and RA2: *Journal of the National Cancer Institute*) and I then look at the initial sentences from journalistic accounts which refer to this research as a 'breakthrough' (JA1: *The Daily Telegraph,* JA2: *The Independent,* JA3: *The Guardian* and JA4: *The Birmingham Post*). In each case, I attempt to find the sequence in two corpora: the British National Corpus (BNC) for the general language and the Pharmaceutical Sciences Corpus (PSC) for scientific discourse. For each sequence, it is possible to multi-word searches of the form *the * of* * (although AntConc often interprets the * symbol as more than one word). When a sequence turns out to be impossible to find (which is the case for most examples above approximately 10 signs), I then search for increasingly shorter extracts.

(RA1) *Trends in Pharmaceutical Sciences*[6]

The first sentence of this research article reads as follows:

**(1)**
Progressive weight loss is a characteristic feature of malignant diseases, and some studies suggest that nearly 90% of patients are affected. (RA1)

Here is the sequence of grammatical items used in the search:

* * * is a * *of * *s, and some *s * that *ly * of *s are *ed. (RA1)

I have included the sign *is* in the search sequence, even though it is used here as a lexical, copula verb (the verb *are* in the second half of the

extract is an auxiliary, a more *bona fide* grammatical item). One justification for doing this is that if *is* is not included, a search for the sequence * * * * *a* * **of* produces too many hits and includes many irrelevant patterns. Using *is* to narrow down the search, I find 264 examples of the sequence: *is a * * of *, and* in the BNC. Many of these examples do not include a clause break before *and,* as in *this show is a triumphant affirmation of life and vitality*, and very few involve a complex nominal (as we have at the beginning of extract 1). Only 11 BNC examples are structurally close (but still not exactly matching) extract 1. However, it is interesting to note how similar these examples are topically to extract 1, all involving highly technical subject nouns and an attributive clause which either defines or evaluates the subject as a more general "cause", "source", "method", "product" etc.:

**(1a)**
COPD **is a leading cause of** morbidity and mortality worldwide, and results in an economic and social burden that is both substantial and increasing. (BNC)

**(1b)**
Pamidronate, a second-generation bisphosphonate, **is a potent inhibitor of** resorption, and has been successful in the treatment of TIH. (BNC)

**(1c)**
This dividing technique **is a useful method of** increase, and works well, provided each piece has some root and some dormant buds or young shoots. (BNC)

A similar search in the PSC of course finds RA, plus 10 other examples, this time with complex nominal structures in subject position. In Gledhill (2000) I found that the sequence *is a* is a salient sequence in research article introductions. It seems that this usage is simply one realisation of a more general pattern in academic writing, in which *to be* introduces an attributive complement in the present tense and has the discourse function of expressing explicit evaluation. As can be seen in the following PSC examples, the phraseology of these patterns is similar to that of the BNC, except that the complement typically refers either to a key biochemical agent / participant, or to a source / cause from the point of view of the observer (example 1f):

**(1d)**
The present inhibition studies show that MAMC **is a competitive inhibitor of** dextromethorphan, and vice versa. (PSC)

**(1e)**
The oncogenic Bcr-Abl tyrosine kinase **is a potent inhibitor of** apoptosis, and it is retained exclusively in the cytoplasm of transformed cells. (PSC)

**(1f)**
This reliance on symptomatic presentation and recall of poorly defined symptoms **is a significant source of** bias and results in underestimates of the true incidence of each event. (PSC)

The second half of extract 1 involves a reporting clause (a projecting clause, in systemic functional terms: Halliday and Matthiessen 2004). Although it is not possible to find the precise sequence *and some *s * that *ly * of *s are *ed* in the BNC, it is possible to find the first part of the structure *and some *s * that* (27 examples). Most of these correspond to a projecting clause with a fairly predictable set of subjects (*analysts, accounts, commentators, estimates, reports*) and verbs (*argue, believe, claim, indicate, suggest*). The PSC does not contain any clause of this type introduced by *and*, but does have 20 examples of projecting clauses of the form *some *s * that* (to cite just one example *While some authors recognize that acute post-operative airway obstruction is common...*). The second half of this sequence (the projected clause *that *ly * of *s are *ed*) is more problematic: the only matching sequence which can be found in the BNC is **ly * of *s are *ed*, and none of these examples have the same grammatical structure (except for some marginally related examples, such as *eventually sets of compounds are perceived... only representatives of parties are elected*). A search of the PSC however produces 82 hits of the form *modal Adv A N of N*, although few of these occur as a projected clause. It is also notable that while the sequence **ly * *s of* matches almost exactly the statistical analysis of clinical methods or the reporting of results that we have in extract 1, the verb forms used are more likely to be past active or past passive:

**(1g)**
**Approximately one third of respondents were** within 10% of the Australian rate (PSC)

**(1h)**
it could be expected that **significantly lower doses of antioxidants could be used** in future prevention studies (PSC)

**(1i)**
Hence, **significantly greater levels of eosinophils were recovered** from the lungs of antigen-challenged mice following prior treatment (PSC)

Thus the only major difference between these examples and extract 1, is that our original sentence uses the present tense, which is more usual in Introductions. It is tempting to suggest that the originality of extract 1 (in contrast to the PSC, and perhaps the general discourse of science writing) is that it adopts a canonical introductory style initially, but then shifts into the phraseology of results reporting. This sort of shift must surely be related to what follows in the argumentation structure of this article. Of course this comment highlights the limits of analysing sentences in complete isolation from the rest if their original context.

(RA2) *Journal of the National Cancer Institute*[7]

The first sentence in this research article reads as follows:

**(2)**
Recently, considerable attention has been directed toward the isolation and identification of the factors responsible for the complex metabolic changes associated with cancer cachexia. (RA2)

Here is the same sequence as a discontinuous framework:

*ly, * * has been *ed toward the * and * of the *s * for the ***s* ed with * * (RA2)

The first few signs in this sequence **ly, * ** provide too many hits in the BNC. But a direct search for *has been *ed toward* does not give any results. However, when I change the verb form (*are / was / were*) and the preposition (*towards,* which is more typical of British English), I find a large number of structurally similar examples (approx. 250). The verbs occurring in this pattern have a consistent meaning (*aimed, directed, orientated*), and the subject nouns refer consistently to research-oriented processes or general cognitive processes (*activity, effort, study*):

**(2a)**
At Sunbury, XTP's **activities are directed towards** helping the business add value (BNC)

**(2b)**
most of these **efforts were directed towards** reducing non-oil imports, which had damaging effects on domestic production. (BNC)

**(2c)**
Early evaluation **studies were directed towards** the use of specific media (BNC)

The same search of the PSC reveals 18 examples of *(be) *ed toward* and 14 examples with the spelling *(be) *ed towards.* As with the BNC data, no examples in the present perfect can be found. The general pattern relates to research activity (around half of the time involving the verb *directed*, the same pattern as that of extract 2 (examples 2a-c). A variation on this pattern (examples 2d-f) involves more observation-oriented verbs (*shifted, skewed, weighted*), relating to the changing 'shapes' of empirical data:

**(2d)**
Future **research** in this area should be **directed toward** tracking the intracellular signal transduction pathways (PSC)

**(2e)**
Particular **interest** was **directed towards** a careful topographical analysis of the obtained proliferation data within different sites of the vessel wall. (PSC)

**(2f)**
The maximal luciferase activity was the same for the three steroids, but the **curve** obtained with testosterone was **shifted toward** a higher concentration of ligand. (PSC)

The second regular pattern to be observed in extract 2 involves the sequence: *the *s * for the ***s*. The BNC has 56 examples of this sequence, although many of these are post-modified nominal groups involving a past participle (of the type: *tested for*). A small handful of examples resemble extract 2 more closely, with a post-modifying epithet such as *accountable, available, responsible:*

**(2g)**
the persons **accountable for** the duty in terms of Section 44 of the Finance Act (BNC)

**(2h)**
the assets **available for** the floating charge holders. (BNC)

**(2i)**
the organisms **responsible for** the sexually transmitted diseases (BNC)

The PSC has one example of sequence *the *s * for the * * *s* and 29 examples of the shortened sequence *the *s * for the * *s.* As in the BNC, most of these are embedded past participle clauses such as *The coefficients obtained for the 8 reactions.* A further set is built around nouns such as

*gene, sequence* post-modified by an embedded progressive participle *N + coding for*. The final pattern (6 examples) involves *responsible for* which, as we saw in the BNC examples, post-modifies a key participant (*agent, enzyme, factor, mechanism*) and introduces a biochemical process*:*

**(2j)**
the composition and nature of the **agents responsible for** the early metasomatic events. (PSC)

**(2k)**
The **enzymes responsible for** the synthesis of gramicidin S suffer degradation or inactivation. (PSC)

**(2l)**
Although the **mechanisms responsible for** the described effects require(s) additional studies...(PSC)

It is striking that in both patterns observed in extract 2 (the verbal group 2a-f and the nominal group 2g-l), only a very restricted set of lexical items are involved in the most closely matching sequences, in fact the same as the ones in extract 2: *directed toward(s) / responsible for*. In both cases we are dealing with conventional lexico-grammatical patterns in academic discourse: one a dynamic metaphor of topicality, roughly equivalent to 'X is of interest to': *attention* / e*ffort / interest ... is / has been directed towards*, the other a stative expression of causality, equivalent to 'X is the cause of Y': *agents / factors / mechanisms ... are responsible for*. The originality of extract 2 is that it exploits both of these phraseologies simultaneously, embedding the *responsible for* pattern within a topical introduction built around *directed toward(s).*

(JA1) *Daily Telegraph*[8]

We now turn to the analysis of initial sentences in journalistic accounts (JA) of cancer research. The first sentence of JA1reads as follows:

**(3)**
Scientists are a step closer to developing an early detection test and possible treatment of four of the most common and intractable cancers. (JA1)

Here is the sequence of grammatical items used in the search:

*s * a * *er to *ing an * * * and * * of * of the most * and * *s. (JA1)

The BNC has 16 examples of the sequence **sa * *er to *ing*. This corresponds to a very predictable pattern involving a verb group expressing movement or proximity (*being* / *bringing* / *taking* + *a step closer* or *a little / one step nearer*) and a complement relating to a scientific discovery (a nominalised mental process *finding, understanding*). The explicit reference to scientists in most of the examples points to journalism rather than academic science:

**(3a)**
SCIENTISTS believe they are **a step nearer to finding** the cause of Cot Death Syndrome, (BNC)

**(3b)**
but scientists are **a bit nearer to understanding** what goes on at the molecular level (BNC)

**(3c)**
Three papers published recently in Science move us **a little closer to understanding** the basis of the disease (BNC)

These are clearly instances of a very regular lexico-grammatical pattern used in journalism. No examples of this pattern can be found in the PSC (although some similar structures can be found, they are unrelated).

Conversely, the second part of JA1 (and in particular the sequence *of the most * and*) appears to be closer to the elaborate nominal constructions of academic writing. The PSC contains 5 examples of the sequence *of * of the most * and*. All of these are instances of the same pattern: a complex nominal in which a biochemical product or process is post-modified by a superlative epithet introduced by *most* and emphasised (sometimes redundantly) by a second epithet introduced by *and.* As can be seen, writers in the PSC have a preference for *effective* + epithet:

**(3d)**
a determination **of the most effective and efficient** biomarkers (PSC)

**(3e)**
Overview **of the most effective and convenient** reagents, (PSC)

(**3f**) Acyclovir is one **of the most effective and selective** agents against herpes viruses (PSC)

(JA2) *The Independent*[9]

The first sentence of extract 4 reads as follows:

**(4)**
A substance found in fish oil is to be used in the treatment of cancer, after new evidence that it can shrink solid tumours and may halt the dramatic weight loss associated with the disease. (JA2)

A * *d in * * is to be *ed in the * of *, after * * that it can * * *s and may * the * * * *ed with the *. (JA2)

Here I shall only concentrate on the first main clause and the sequence: *A * * *d in * * is to be *ed in.* Although it is difficult to find the exact same sequence in the BNC, especially with three contiguous lexical items, over 50 similar examples of *a/n * found in N* can be found. This is an instance of a very regular pattern involving a complex nominal referring to a *body* or *substance,* post-modified by an embedded clause expressing a cognitive process of discovery (*encountered, found, identified*). These correspond to two related patterns: a) the finding of a body in a journalistic account, and more commonly b) a scientific 'finding' involving a specific substance. Here are examples of both from the BNC:

**(4a)**
**The body of a boy found in the River Thames is to be** re-examined to see if he was the victim of a ritual killing. (BNC)

**(4b)**
**A potentially harmful chemical commonly found in plastic baby bottles is to be** banned from their manufacture from next year. (BNC)

**(4c)**
**A substance found in yew trees** may help cancer sufferers, reports Carina Norris (BNC)

It might be thought that this pattern would also be common in science writing. However although the same structure is found in the PSC (32 examples), the pattern is not quite the same. In the PSC, the preposition *in* tends to be used to introduce nominalised processes rather than locations (as mentioned in section 2.3, in Abstracts the pattern is *found / observed in the [process X] of [participant /product Y]...*). When locations are referred to in the PSC, they are introduced by a variety of verbs expressing specific material rather than general cognitive processes, and in contexts where the properties of biochemical products are defined as locations or roles (expressed after *in*):

**(4d)**
With fibronectin, **a glycoprotein deposited in** the basal membrane after debridement that produce (PSC)

**(4e)**
The present study demonstrates for the first time that MIF, **a cytokine involved in** the inflammatory process, is produced by human trophoblasts (PSC)

**(4f)**
Finally, BAL fluid contains significant levels of IL-16, **a cytokine implicated in T** lymphocyte recruitment (PSC)

Returning to extract 4, it is notable that both the extract and two BNC examples (4a-b, above) make use of the "infinitival future" (*is to be used*). I would suggest that form (combined with the passive) is more akin to the discourse of research articles than journalistic accounts. If we look in the BNC for sequences such as *is to be *ed in the * of* (the main clause in extract 4), we find examples such as the following (4g-i). It is also notable that in each case the processes expressed are closer to those of extract 4 and the PSC (*employed, implemented, used*):

**(4g)**
its potential **is to be employed in the evaluation of** patients with RA consistently, with close frequency, and independently of any calculating device … (BNC)

**(4h)**
A commercial that **is to be implemented in the teaching of** a textbook unit (BNC)

**(4i)**
The equipment provided in support of the AED Program **is to be used in the event of** an SCA at Gonzaga University. (BNC)

It would seem then that extract 4 is another example of a relatively "hybrid" style, with at least two lexico-grammatical patterns belonging to different types of discourse. The first main clause employs a prototypically journalistic way of presenting a "finding" (nominal post-modification of *substance*), although the core predicate in the clause expresses the "future" in a prototypically impersonal, academic manner (*is to be used*). Although space precludes further analysis here, it is notable that the rest of the sentence involves structures which are typical of elaborate scientific prose (the embedded nominal projection *after new evidence that...* post-

modification after *associated with,* etc.), but also typical of the phraseology of journalism (notably the choice of epithets: *new evidence, solid tumours, dramatic weight loss*).

(JA3) *TheGuardian*[10]

The first sentence of extract 5 reads as follows:

**(5)**
A substance found only in oily fish may help to fight one of the main symptoms of cancer as well as leading to new forms of treatment for some of the most resistant tumours. (JA3)

Here is the same sequence stripped of its lexical items:

A * *d only in * * may * to * one of the * *s of * as well as *ing to * *s of * for some of the most * *s. (JA3)

As with previous extracts, there is only space here to analyse one or two key features of this sentence. Perhaps the most important structure here is the post-modified subject *A * *d only in...,* which involves the same *substance found in* pattern as in extract JA3. The only difference is that the substance is found *only* in one source or location. The following BNC examples give a flavour of the general pattern:

**(5a)**
Studies suggest that **a substance found only in** types of manuka honey may help prevent plaque from damaging calcium phosphate in tooth ... (BNC)

**(5b)**
An American study has found that theaflavin-2 - **a substance found only in** black and oolong teas - was able to induce aptosis (cell death) in ... (BNC)

**(5c)**
The cause of this common neurological disease is thought to be a dietary factor or toxic **substance found only in** that area. (BNC)

As mentioned above, this pattern occurs in the PSC, but in a somewhat different form: the closest examples are active and passive finite clauses such as *viral antigens were found only in the bronchial epithelium.* Also, verbs other than *found* are more likely to be used, most notably *identified* and *observed.*

The second key element of phraseology in extract JA3 is a verbal group complex introduced by *may *to * one of the * *s of*. There are no exact matches for this pattern in the BNC, although it is possible to find variations, most notably with different modal verbs or other morphological changes (such as *help + to V to / help V-ing*) or different superlatives (*one of the main / most*). The examples cited below belong to a variety of relatively formal expository genres, but it is notable that 5e and 5f are also clear examples of introductions to a "breakthrough" story. All examples share the same discourse function: the subject / theme of the clause is the key to solving or understanding some problem:

**(5d)**
It may **help explain one of the underlying causes** of coral decline, and is one of the most comprehensive analyses yet done on the types of viruses in a … (BNC)

**(5e)**
New research from the University of Adelaide **could help protect one of the world's most globally threatened tree species** - the big leaf mahogany - from ... (BNC)

**(5f)**
TAKE two aspirin and have a lie down used to be a doctor's cliché, but a study says it could **help fight one of the world's biggest killers** (BNC)

This phraseology is not reflected in the PSC, except for unrelated structures (*the reaction sequences ... may happen to give one or more of the products indicated*). Verb groups like *help (to) explain* and noun groups like *one of the main symptoms* can both be found in the PSC, but neither are used in the same context. The following examples from the PSC give an idea of the more usual phraseology of *may help + V* as a statement of research goals (5g-h) and the use of the superlative in defining clauses (5i-j):

**(5g)**
In general, genotyping **may help to categorize** the patients as mild HPA at an early stage of life. (PSC)

**(5h)**
The study of gene-environmental interactions **may help enhance** our understanding of how these factors cause cancer formation. (PSC)

**(5i)**
The release of fluoride **is one of the main advantages of** glass ionomer cements. (PSC)

**(5j)**
Nausea **is one of the primary symptoms** in anxiety disorders and the effect of angina on nausea may be an indirect effect via anxiety. (PSC)

(JA4) *Birmingham Post*[11]

The first sentence of extract 6 reads as follows:

**(6)**
Birmingham scientists believe they are on the verge of beating cancer. (JA4)

* *s * they * on the * of *ing *.(JA4)

Unlike the previous extracts (3, 4 and to a lesser extent 5), in which I find aspects of both journalistic and scientific discourse, extract 6 is almost purely journalistic. Phrases such as *scientists believe* and *beating cancer* belong prototypically to the dramatic language of a breakthrough story. Such phrases are absent of course from the PSC, where the verb *believe* is either expressed in the passive or with *we* as subject, and *beat* is a noun (*heart beat*). However, the sequence *on the * of *ing* is worth analysing in more detail. A search for this sequence results in 102 hits in the BNC, and involves a very predictable but also productive pattern of the form *(to be) on the (brink, eve, path, point, verge) of V-ing*. The discourse function of this sequence is similar to the more academic infinitival future (*is to be*) we saw in extract 5: a news story is in the process of breaking (emphasis on the "here and now") according to some higher authority or source (often expressed in a projecting clause *Scientists believe that...*) The BNC examples suggest a broad set of technical or (pseudo-)scientific topics for the breakthrough:

**(6a)**
He demonstrates that municipalities are **on the brink of learning** how to rezone and use other land use and devel- opment techniques that significantly reduce carbon. (BNC)

**(6b)**
"We are **on the eve of settling** the deal" the official, who is close to talks with Israel, told Reuters. (BNC)

**(6c)**
Many scientists believe that we are **on the verge of contacting** alien life-forms. (BNC)

In systemic functional terms (Halliday and Mathiessen 2004), this structure corresponds to a verbal group complex such as "to go on doing", "to keep doing", in which an initial verb (or in this case a prepositional group) expresses the "phase" or aspect of the following verb. Phase is sometimes encountered in scientific articles (*one patient, though, did go on to develop persistent ventricular arrythmias, Pain continues to be one of the more frequent causes of unplanned admission*) but this is not common, and no examples of the sequence *on the * of *ing* correspond to this pattern in the PSC (except for unrelated sequences such as *our study focused on the consequences of identifying...*).

## 4. Discussion

The following table sets out the basic results of the survey carried out in this study:

| Extract | Patterns typical of science writing | Patterns typical of journalistic writing |
|---|---|---|
| RA1 | ... is a characteristic feature of...<br>... some studies suggest that nearly 90% of... | |
| RA2 | ... attention has been directed toward ...<br>... responsible for the complex metabolic changes ... | |
| JA1 | ... of four of the most common and intractable... | Scientists are a step closer to developing ... |
| JA2 | … is to be used in the treatment of... | A substance found in ... |
| JA3 | | A substance found only in ....<br>may help to fight one of the main symptoms ... |
| JA4 | | Birmingham scientists believe ...<br>... are on the verge of beating ... |

In the analysis I have presented above, I have only been able to point out one or two sequences within each extract which I believe to be "prototypical" in either science writing or journalism. Notwithstanding the

limitations of this analysis, I believe that this initial survey does show that it is possible to use grammatical items to identify extended lexico-grammatical patterns. Generally speaking, each LG pattern is exclusive to one discourse type. For example, the patterns *to be a N closer to V-ing* or *to be on the N of V-ing* in JA1 and JA4 correspond to very productive constructions which express "phase" in the verbal group. Both patterns can be seen to have very specialised discourse functions: they are both used to "break" impending news in journalistic accounts of science. In contrast, the patterns *is a Adj N of* in RA1... and *N has been directed toward(s)* in RA2 are verbal group complexes which are more typical of academic and scientific writing, and they also have their own particular discourse functions (evaluation, topicalisation etc.). It is true that some LG patterns, such as the embedded clause *a N found in* in JA2 and JA3 are found in both types of writing. However, when we look at the extended contexts of this pattern, two sub-patterns emerge: one an expression of definition in academic / science writing (*a [substance, result] [found, observed] in the [treatment] of [disease]*) and the other relating to a discovery in journalism (*a [substance] [found] (only) in [fish oil]*). This picture is complicated by the fact that in some cases (JA2 and JA3), two patterns belonging to different types of discourse can be used in the same extract. In other cases it is possible to observe a shift in discourse patterning within the same general register (RA1's move from introductory style to the present-tense reporting of results). Such instances of hybridity clearly show how variation within a single sentence is determined to a very large extent by the rhetorical functions of the surrounding text. But this does not distract from the general observation that each lexico-grammatical pattern has a distinct discourse function, and that each pattern can be broadly associated with the discourse of research or the discourse of journalism.

What are we to make of the fact that, in the majority of cases, it is in fact rather difficult to find exact matches for sequences of signs (most of the sequences analysed in the previous section involve 10 signs or less, including wild-cards and grammatical morphemes)? In lexicometrics and forensic linguistics, it has long been known that above a certain length, no two sequences of words are ever identical (for example Olsson 2004 sets this limit as low as seven identical words), and that if two long sequences do match, then there must be some degree of mutual influence (plagiarism, crypto-citation etc.). The results set out in this study appear to confirm this view. However, there are one or two factors which complicate this picture. In particular, the viewpoint adopted in this study is somewhat different to that of the forensic linguist. Because I adopt a phraseological perspective, I have no expectation that any stretch of text will be entirely original. It is

an article of faith among the so-called "contextualists" (i.e. Firth, Sinclair, Hunston and others) that the "idiom principle" guides our thoughts and words, and that much of what we can observe in discourse is based on variations of pre-established, predictable patterns of language. So the fact that many exact matches of the sequences I am looking for cannot be found is on the face of things rather surprising. However, it occurs to me that those of us who work on collocations, "fixed expressions" and other phraseological phenomena often forget the inherent creativity and variability of what we have come to think of as "formulaic" language. One reason for this complexity must be that, especially when we look at lexico-grammatical patterns above the level of the group, we are confronted with complexities at the level of the text. At this level, as we have seen in a number of examples in this study, there is often much variation in morphology and determiner use, features which in English are highly variable and sensitive to factors such as textual cohesion. The converse side of this complexity is that whenever we see even a small stretch of words from any particular text, we are almost always able to predict the text type to which it belongs. If we are able to do this fairly systematically (as shown in cloze tests, and as I have attempted to show here), then the repertoire of patterns which language users are familiar with must be very rich, and our capacity to detect and interpret such a variety of patterns must be very impressive indeed.

## Notes

[1] I am thinking here of the Encode project, which aims to identify and analyse "junk DNA", or what they term "non-coding functional elements" in the human genome (Maher 2012).

[2] In the original PSC there were 150 RAs, with a total of approx. 500,000 words. Since then, I have added 250 articles to bring to total to over 1.8 million.

[3] For these data, I used the Log Likelihood metric calculated by AntConc version 3.2.4. My initial results in Gledhill (1995) were obtained using Chi-squared. Generally, Chi-squared gives more weight (= higher keyword scores) to high-frequency items, i.e. grammatical words. However, I have conducted this test using both measures, and as far as I can see, with Chi-squared, the same items occur with different scores but in the same relative order.

[4] This figure is reproduced from Gledhill (1995, 30).

[5] The reason for choosing these texts is that in 1992, while I was compiling the PSC, the research carried out by the Pharmaceutical Sciences team (under Michael Tisdale) was reported as a 'cancer breakthrough' story in over a dozen articles in the local and national press in the UK. I have included a reference to each text used in the following sub-sections. Unfortunately, I have not been able to identify the authors of all of the journalistic texts.

[6] Tisdale, Michael. 1990. Newly identified factors that alter host metabolism in cancer cachexia. *Trends in Pharmaceutical Sciences.* 11(12): 473-475.
[7] Beck, S.A. Mulligan, H., Tisdale, M. 1990. Lipolytic factors associated with murine and human cancer cachexia. *Journal of the National Cancer Institute* 82: 1922-1926.
[8] "Cancer discovery by farmer scientis". *Daily Telegraph* 28 November 1992.
[9] Hunt, Liz. 1992. "Chemical in fish oil to be used to treat cancer". *The Independent* 30 December 1992.
[10] "Fish acid may help cancer victims". *The Guardian*, date not recorded.
[11] "Midland team may have cancer cure within year". *Birmingham Post,* 30 July 1992.

## References

Anthony, Laurence. 2002. *A machine learning system for the automatic identification of text structure and application to research article abstracts in computer science*. PhD Thesis, University of Birmingham, Birmingham.

Biber, Douglas, Susan Conrad, Randi Reppen, Pat Byrd, and Marie Helt. 2002. Speaking and writing in the university: A multidimensional comparison. *TESOL Quarterly* 36(1): 9-48.

Biber, Douglas, Susan Conrad, and Viviana Cortes. 2004. "If you look at...": Lexical bundles in university teaching and textbooks. *Applied Linguistics* 25(3): 371-405.

Biber, Douglas, and Federica Barbieri. 2007. Lexical bundles in university spoken and written registers. *English for Specific Purposes* 26: 263-286.

Bordet, Geneviève. 2011. *Étude contrastive derésumés de thèse dans une perspective d'anaIyse de genre*. Thèse de doctorat, 28 avril 2011. Université Paris Diderot/Paris7.

Cheng, Winnie, Chris Greaves, John McH. Sinclair, and Martin Warren. 2008. Uncovering the extent of the phraseological tendency: Towards a systematic analysis of concgrams. *Applied Linguistics* 30(2): 236-252.

Crystal, David. [1991] 2008. *A dictionary of linguistics and phonetics* (6th Edn). London, Blackwell.

de Cock, Sylvie 1998. A recurrent word combination approach to the study of formulae in the speech of native and nonnative speakers of English. *International Journal of Corpus Linguistics* 3(1): 59-80.

Firth, John Rupert. 1957. *Papers in linguistics 1934-1951*. Oxford: Oxford University Press

Gledhill, Christopher. 1995. Collocation and genre analysis. The phraseology of grammatical items in cancer research articles and abstracts. *Zeitschrift fiir Anglistik und Amerikanistik* XLIII1(1): 11-36.

—. 2000a. *Collocations in science writing*. Tübingen: Gunter Narr.

—. 2000b. The discourse function of collocation in research article introductions. *English for Specific Purposes* 19: 115-135.

—. 2011. The lexicogrammar approach to analysing phraseology and collocation in ESP texts. *Anglais de Spécialité* 59: 5-23.

Groom, Nicholas. 2005. Pattern and meaning across genres and disciplines: An exploratory study. *Journal of English for Academic Purposes* 4(3): 257-277.

—. 2010. Closed-class keywords and corpus-driven discourse analysis. In *Keyness in texts*, ed. Marina Bondi and Mike Scott, 59-78. Amsterdam and Philadelphia: John Benjarnins.

Halliday, Michael, and Christian Matthiessen. 2004. *An introduction to functional grammar* (3rd Edition). London: Arnold.

Hasselgren, Angela. 2002. Learner corpora and language testing: Small words as markers of learner fluency. In *Computer learner corpora, second language acquisition and foreign language teaching,* ed. Sylviane Granger, Joseph Hung and Stephanie Petch-Tyson, 143-173. Amsterdam: John Benjamins.

Hoey, Michael. 2005. *Lexical priming: A new theory of words and language*. London: Routledge.

Hunston, Susan. 2008. Starting with the small words: Patterns, lexis and semantic sequences. *International Journal of Corpus Linguistics* 13: 271-295.

Hunston, Susan, and Gill Francis. 1998. Verbs observed: A corpus-driven pedagogic grammar. *Applied Linguistics* 19(1): 45-72.

—. 2000. *Pattern grammar*. Amsterdam: John Benjamins.

Hyland, Ken. 2008. As can be seen: Lexical bundles and disciplinary variation. *English for Specific Purposes* 27: 4-21.

Lee, David Y.W., and Chen Xiao. 2008. Small words, big deal: Teaching the use of function words and other key items in research writing. In *Proceedings of the 8th Teaching and Language Corpora Conference*, ed. Ana Frankenberg-Garcia, Tawfiq Rkibi, Maria do Rosário Braga da Cruz, Ricardo Carvalho, Direito Cristina and Diogo Santos-Rosa, 198-206. Lisbon: Associação de Estudos e de Investigação Científica do ISLA.

Luzón Marco, María José. 1999. The phraseology and meanings of the pattern be+adjective + to-infinitive. *La Linguistique* 35(2): 47-60.

—. 2000. Collocational frameworks in medical research papers: A genre-based study. *English for Specific Purposes* 19(1): 63-86.
Maher, Brendan. 2012. Encode: The human encyclopaedia. *Nature* 489 (7414): 46-48.
Meyer, Paul. 1988. Statistical text analysis of abstracts: A pilot study on cohesion and schematicity. *Computer Corpora Des Englishen* 3: 17-40.
Olsson, John. 2004. *An introduction to language, crime and the law.* London: Continuum Books.
Renouf, Antoinette. 1992. "What do you think of that?" A Pilot study of the phraseology of the core words of English. In *New directions in English language corpora*, ed. Gerhard Leitner, 301-317. Mouton de Gruyter: Berlin.
Renouf, Antoinette, and John McH. Sinclair. 1991. Collocational frameworks in English. In *English corpus linguistics*, ed. Karin Aijmer and Bengt Altenberg, 128-143. London: Longman.
Riloff, Ellen. 1995. Little words can make a big difference for text classification. In *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval,* July 9-13 1995, 130-136. Seattle, Washington.
Scott, Mike, and Chris Tribble. 2006. *Textual patterns: Keyword and corpus analysis in language education.* Amsterdam: John Benjamins.
Simpson, Rita. 2004. Stylistic features of academic speech: The role of formulaic expressions. In *Discourse in the professions: Perspectives from corpus linguistics*, ed. Ulla Connor and Thomas A. Upton, 37-64. Amsterdam and Philadelphia: John Benjamins.
Sinclair, John McH. 1991. *Corpus, concordance, collocation*. Oxford: Oxford University Press.
—. 1995. *Collins Cobuild English dictionary*. (2nd Edition). London: Harper Collins.
Smadja, Frank. 1993. Retrieving collocations from text: Xtract. *Computational Linguistics* 19(1): 143-177.
van der Wouden, Teun. 2001. Collocational behaviour in non-content words. In *Collocation: Computational extraction, analysis and exploitation. Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics and the 10th Conference of the European Chapter*, ed. Béatrice Dalle and Geoffrey Williams, 16-23.
—. 2007. On the phraseology of stop words. *Leiden Papers in Linguistics* 4(1): 56-67.
Vergne, Jacques. 2004. Découverte locale des mots vides dans des corpus bruts de langues inconnues, sans aucune ressource. In *Le poids des*

*mots, Actes des 7es Journées Internationales d'Analyse Statistique des Données Textuelles*, ed. Gérald Purnelle, Cédrick Fairon and Anne Dister, 1158-1165. Louvain-la-Neuve 10-12 mars 2004 / March 10-12, 2004 (JADT vol. 2).

Williams, Geoffrey C. 1998. Collocational networks: Interlocking patterns of lexis in a corpus of plant biology research articles. *International Journal of Corpus Linguistics* 3(1): 151-171.

# CHAPTER TWO

# THE ROLE OF "LEXICAL PAVING" IN BUILDING A TEXT ACCORDING TO THE REQUIREMENTS OF A TARGET GENRE

GENEVIÈVE BORDET
UNIVERSITÉ DE PARIS DIDEROT, FRANCE

## 1. Introduction

The lexical combinations I explore in this study are lexico-grammatical patterns that appear across a text, at specific sections, to convey and variously articulate the same key concept. As such, their analysis calls for knowledge and methods developed in the fields of lexicogrammar, genre studies and discourse analysis. Lexicogrammar and genre-based discourse studies have drawn the attention of both discourse analysts and corpus linguists over the last thirty years (e.g. Renouf and Sinclair 1991; Gledhill 1995; Hoey 2005). This study aims to explore connections between genre and text analysis with the study of variations in lexical combinations around pivot keywords. Interest in genre studies is largely due to Swales' (1990) well-known approach, which considers genre as the result of the interaction between a context, a discourse community and a communicative purpose: these three factors determine the type of text and the type of communicative strategy. His approach offered valuable insight into the external criteria which make a text part of a genre.

However, we are left with one key problem: is it possible to describe the internal criteria which define a genre, i.e. which linguistic features are typical of a genre and mark the text as belonging to a specific genre? The most successful recent exploration of this aspect is based on the identification of rhetorical moves inside specialised discourses (e.g. Biber et al. 2007). Other studies have deepened the connection between the move structure in a text and the genre it belongs to. They have focused more specifically on the connection between disciplinary values and

realization of the genre (Bondi 2002; Hyland 2002; Charles 2003). Another approach has highlighted the interaction between genre, moves and lexicogrammatical patterns, referring to the complex associations of words. One case in point is Gledhill's demonstration as to how a specific succession of lexicogrammatical patterns builds the medical discourse on cancer (Gledhill 1995). This is where the issue of lexicogrammatical patterns' textualising role impacts on the description of genre.

The discourse function of regular lexicogrammatical patterns, including collocations, has been a major focus of interest since Firth (1957) first prioritized "contextual meaning" over "conceptual meaning". Corpus linguistics and the development of technological tools have increased that interest (Renouf and Sinclair 1991; Teubert 2007), since they offer a new insight, based on wide corpus analysis, into regular lexicogrammatical patterns as characteristic of specific types of discourse. Besides this statistical approach, the cohesive role of collocations at text level has also long attracted attention (Halliday and Hasan 1976). More recently, Hoey (2005) and Gledhill (2009) have highlighted the textualising role of regular lexicogrammatical patterns and collocations, calling for a contextualized approach. "Textual collocation" builds the linearity of text seen as a process rather than a product (Partington 1998, 15). To study this building process, it is necessary to reconcile the top-down statistical approach of corpus linguistics, from corpus to text, and the bottom-up discourse analysis approach from text to corpus (Biber et al. 2007). While the text-structuring role of collocations is rarely denied, their fuzzy nature contributes to making their identification and functional analysis difficult. Collocations, co-occurrences, and colligations all point to repeated lexicogrammatical combinations which occupy an intermediary position on the cline from the fixed multiword units to freely combined multiword associations. Taking into account the variations of this pervasive linguistic phenomenon, so as to assess its text-structuring role, leads to various types of approach. The focus is alternatively set on the grammatical and the lexical aspects of collocations. Hoey (2005) proposes that each word is "primed" by its collocational use to take a specific meaning in a specific communicative context, and studies the grammatical use of lexicon as a result of this initial priming through use in context. Gledhill (2011) points out the fact that "lexical items that are usually involved in cohesive chains are necessarily embedded in lexicogrammatical patterns, whose distribution throughout the text must therefore contribute to the development of coherence throughout the text" (Gledhill 2011, 13). On the lexical end of the cline, Bondi (2010, 4) investigates the structures of textuality through the reiteration of keywords and "the patterns created in

text between their collocates". She claims that "the key lexical elements of a text create a dense network of intercollocations including both continuous and discontinuous phraseological patterns" (Bondi 2010, 4).

The point of my research is to focus on the text-structuring role of the reiteration of key pivot terms throughout the text and the variation of lexicogrammatical patterns in which a given pivot term can be observed. "Pivot terms" are terms that are reiterated along the text, therefore focusing the reader's attention on these terms and their collocates. This type of lexical chaining is considered as a potential cohesive device at text level. It is also assumed that this cohesive effect contributes to the specific communicative aims of the genre.

## 2. "Lexical paving" as a text-structuring and a generic device

Within the context of the text-structuring role of regular lexicogrammatical patterns and considering the way they set the text inside a genre (Gledhill 2009), I intend to focus on the reiteration of key lexemes and the evolution of their lexical environment within a text, seen as a specific feature which can contribute to the realization of the genre.

Here the focus is on reiterated co-occurrences of pivot terms within the text and their distribution along the rhetorical structure. These pivot terms are considered as potential keywords insofar as they attract the reader's attention to key concepts repeated along the text, creating thus a sort of echo or an intratextual lexical chain based on isotopy. I argue that a succession of lexical patterns' variations around reiterated pivot keywords within a text forms a sort of "lexical paving"[1] whose interaction with the rhetorical moves contributes to the coherence of the argumentation in a text, as expected by a specific discourse community.

My hypothesis is that this rhetorical device is a powerful tool in making the text part of a genre i.e. rendering it adequate to the communicative characteristics of the discourse community in point. This study aims to verify that lexical variation around recurrent pivot terms not only reveals individual strategies but is also typical of a generic structure.

Based on Biber et al.'s (2007) "bottom-up and top-down" approach, abstracts from two disciplines have been collected and studied both at text and corpus level. First, each text has been marked for rhetorical moves and reiterated terms have been identified, starting from the text title. These terms have been listed so as to give a picture of the ontology of the discourse for each discipline (Bondi 2010, 8) (see Appendix 1). Then, lexical variations around these pivot terms have been analysed, with

regard to their interaction with the rhetorical moves, in an attempt to ascertain their "textual-pragmatic meaning" (Bondi 2010, 4). The point is to understand whether, beyond the simple reiteration of pivot terms inside the text and the corpus, the variations of their lexical environment, at text level, contribute to shifting the focus from one move to the next, therefore reinforcing the persuasive effect of the genre.

## 3. Abstracts as a genre

Academic abstracts have been considered as an interesting genre for this type of research for several reasons. Although they have long been considered a "sub-genre" (Swales and Feak 2009), in the context of a dramatic increase in scientific publications, abstracts have become the major "gatekeepers" (Swales 1990) of the academic field. Researchers cannot afford to read whole texts unless they can reasonably expect them to meet their informational needs; and the abstract provides an opportunity to judge the content of a paper. Moreover, the spectacular increase of subscription prices in a field dominated by a small number of publishers makes access to scientific information paradoxically more and more difficult (Bordet 2014). Abstracts databases[2] are therefore widely used as a "screening device" by researchers so as to identify relevant literature since they function as "stand-alone mini-texts", which often appear separately from the paper (Huckin 2001). They also offer free-access when access to papers themselves is most of the time restricted to subscriptions.

A further point is that within the context of increased competitiveness ("publish or perish"), abstracts have become writers' "self-promotional tools" (Hyland 2000). As such they can be expected to reflect the targeted communities' expectations:

> Abstracts are worthy of study because they are significant carriers of a discipline's epistemological and social assumptions, and therefore a rich source of interactional features that allow us to see how individuals work to position themselves within their communities. (Hyland 2004, 63)

Finally, abstracts offer a concentrated version of genre-specific discursive strategy, since their writers are only allowed a few words to demonstrate both the coherence of their research and its adequacy as to their readers' expectations.

This study will focus specifically on PhD thesis abstracts, despite the fact that this type of abstract clearly does not currently play the same crucial gatekeeper role as do research article abstracts. They are often written after the completion of the PhD thesis and do not seem to be a

decisive criterion for the validation of the thesis. However, they offer an interesting view of a field of research as seen by newcomers. One may assume that they reflect the image PhD students have acquired, during their research work, of the targeted scientific community in terms of epistemological values and linguistic expectations. As such, they offer an interesting insight into both the image projected by the community and into its perception by "would be insiders" (Hyland 2000). Although there may be variations due to the differences in the academic background of the PhD applicants, this might help indicate which lexicogrammatical and rhetorical features are perceived, by the applicants themselves, as contributing to the admission of a new "academic voice" (Fløttum et al. 2006; Dressen-Hammouda 2008). The choice of reiterated terms, their lexical combinations and their variations may be considered as one of these features.

## 4. Comparable corpora of PhD abstracts

This study is based on two comparable corpora of PhD abstracts taken from two disciplines: Didactics of Mathematics and Materials Science. Contrasting these two disciplines was expected to help assess the extent to which rhetorical strategies are connected with general scientific values on the one hand, and with specific disciplinary values on the other. A comparison of the two disciplinary discursive patterns should highlight common academic features and any specific disciplinary characteristics.

It must be specified here that both of the chosen disciplinary fields call for additional comments. Didactics of Mathematics is the result of the development of mathematics education as a scientific discipline (Biehler et al. 1994). The objective of Didactics of Mathematics is to study the pedagogy of mathematics, or mathematics education. The term "Didactics of Mathematics" was coined in the 70's and is widely recognized internationally. Materials Science encompasses a wide range of interests ranging from nanotechnologies to composite materials, for instance. In this study the focus was set on PhD theses dealing with fracture mechanics, as one specific sub-discipline of Materials Science and in order to guarantee a better homogeneity of corpus. This was not the case with Didactics of Mathematics, which, as a sub-discipline of Mathematics, is centered on Didactics, rather than on specific branches of mathematics. Didactics of Mathematics and Materials Science were chosen so as to contrast respectively soft and hard sciences. Both disciplines can also, to a certain extent, be considered as interdisciplinary, with education and mathematics on the one side, chemistry and physics on the other.

All abstracts are written in English and were selected primarily with the academic search engine Scirus.[3] The PhD theses were defended in English speaking institutions regardless of the nationality of the author. Reflecting the comparative approach of Didactics of Mathematics, abstracts came from various English-speaking countries (USA, New Zealand, Australia). Although the educational systems are different, mathematics didacticians share the same objects of study. Therefore, this diversity should not hinder the analysis. It must be added that this choice was also made out of consideration for corpus size, and availability, since much fewer PhD theses are published in that field than in Materials Science, for instance. Material Science abstracts mainly come from the USA. The corpora include 30 abstracts each. The very limited size of the corpora (12,500 words for Didactics of Mathematics, 11,500 words for Materials Science) was designed so as to allow a bottom-up and top-down approach, from text to corpus and back to text. As stated above, my aim is not to identify statistically recurrent types of discursive lexicogrammatical variations but to assess both at corpus and text level their type and degree of adequacy as to the genre and its communicative objectives.

## 5. A bottom-up and top-down approach: From text to corpus and back to text

The following section describes a 3 step methodology. The first step starts at text level, with a manual analysis of rhetorical moves. In this stage, pivot terms or keywords are identified for each text. In the second step, a concordancer is used to assess the semantic value of each pivot keyword in the disciplinary corpus, based on their frequency and their distribution, across the corpus and throughout each text. The final step is a discourse analysis of each text. Lexical variations around the reiteration of one or several pivot keywords are identified and classified according to the rhetorical structure. The overall objective is to better understand the role played by the variation of their lexical environment in the progression of the text, through the shift from one rhetorical move to the next. It may also provide insight into the semantic value of the reiterated keywords for the discipline.

### 5.1. Identification of moves and reiterated pivot keywords

This first step includes two tasks for each text: marking of the rhetorical moves (cf. Hyland 2004, 66); identifying reiterated pivot keywords (Bondi 2010).

*Task 1: each text is marked for moves*

With a view to gaining a better understanding of the potential interaction between lexical variations in the immediate context of reiterated pivot keywords and rhetorical structure, each text was marked for moves (Swales 1990; Bhatia 1993; Hyland 2000). The study of the 60 abstracts led to modifications of the initially adopted Hyland's classification, which includes five moves i.e. Introduction, Purpose, Method, Product, Conclusion, since the initial manual analysis gave evidence of a structure comprising only 4 moves: Therefore, Bhatia's (1993) four move structure (Purpose-Method-Results-Conclusion) seemed more suitable to the description of our corpus. However, there was little evidence of a systematic distinction between "Results" and "Conclusion", which led to the choice of a single "Results" move, presenting the outcome of the research and its discussion. Conversely, it seemed relevant to add a "contextualizing" move. This move introduces the research statement (or "Purpose" in Bhatia's terms). It actually "maps the research territory" from both an institutional and theoretical point of view (Malavasi and Mazzi 2010, 182). The prototypical move structure as uncovered in the studied corpus can be described as follows:

- contextualizing the research project (Context);
- formulating the research statement (Research statement);
- describing the method (Method);

stating the results and offering their interpretation (Results). Below an example of this rhetorical pattern is outlined

**(1)**
**[Context_beg]** Mathematics, as a subject, is used in various scientific careers as a selection tool. It is regarded as the cornerstone of scientific literacy. However, since learners in South Africa do not perform optimally in mathematics they do not enjoy international recognition. Education renewal is ongoing, and South Africa currently follows an outcomes-based (OBE) approach. The teaching of mathematics cannot be renewed successfully if assessment methods are not regularly adapted to meet new developments in the field. The incorporation of an OBE approach at school level made it necessary to facilitate assessment renewal in tertiary mathematics at the Tshwane University of Technology (TUT). TUT is engaged in a merger of three institutions, which has made the development of new curricula and teaching material essential. Hence, this is a perfect time to introduce assessment renewal.

**[Context_end] [Research Statement beg]** The primary purpose of this thesis is to report on the research study and its results, and to make recommendations for improving the practice. The overarching research hypothesis in this study is that a suitable assessment would probably enhance the effectiveness of a student's learning. **[Research statement_end]**

**[Method_beg]** The research focused on the following questions:

- To what extent are outcome-based strategies effectively and regularly introduced in the teaching of mathematics at TUT?
- Will tertiary mathematics facilitators be prepared to implement outcomes-based strategies at TUT?
- To what extent are outcomes-based strategies in subjects supported by mathematics implemented at TUT?
- How does the ecology of TUT affect the implementation of outcome-based strategies?
- What other factors could influence the level of implementation of OBS at TUT?
- Have any of the mathematics facilitators at TUT received suitable and adequate training in the implementation of outcome-based strategies?
- What are the possible implications of the study for TUT's assessment policy?

Action research was chosen as the research design because it is ideally suited to improving practice. Quantitative and qualitative data were collected through questionnaires, personal interviews, interviews with focus groups, observations, documentation and a reflective diary. **[Method_end]**

**[Results_beg]** The main findings are as follows:

- OBE strategies are not being introduced throughout TUT in the teaching of mathematics.
- Group work and peer assessments are rare occurrences.
- Some lecturers are convinced that new assessment methods would lower the standard of their teaching.
- Uncertainty about the merger and the varying teaching conditions at the different campuses tend to inhibit lecturers, making them less willing to undertake assessment renewals.
- TUT should review its admission criteria.
- The lecturers cited large class groups, a lack of marking assistance and ignorance about OBE as reasons for failing to undertake assessment renewal.

The study prepared respondents for assessment renewal. In the interim, however, TUT has introduced a Policy on Teaching, Learning and Technology, whereby OBE has been selected as the teaching model for TUT. In future, respondents will receive training and guidance in the implementation of OBE. This study has hopefully made a significant contribution to this positive development. **[Results_end]**

The use of the prototypical phrase "the primary purpose of this research is to" marks the beginning of the research statement. Then, the past tense used in "the research focused on" is a clear indication of the beginning of the "method" move. Use of a further prototypical phrase ("The main findings are") marks the start of the "results move". It must be stated that the contextualizing move is much easier to identify in the "Didactics of Mathematics" corpus (DM), as shown above. In the "Materials Science" corpus (MS), contextualization sometimes only appears under the form of theoretical or experimental models' designation, as in the following abstract:

**(2)**

[**Research statement_beg**] This thesis is a detailed investigation of hydraulically-driven fracture propagation in poroelastic rock. Biot's theory of poroelasticity is used to study coupling between rock deformation and fluid flow within its mass (MS corpus). [**Research statement_end]**
[**Method_beg**] The topic is developed as follows: (1) a nonlinear fracture mechanics model is adapted for a poroelastic continuum, (2) poroelastic concepts and effects are illustrated through application to the 1-D, PKN fracture model. (...) An iterative, staggered solution procedure, which implicitly advances the solution at each time step, has been designed to take advantage of vector processing on a mini-supercomputer. [**Method_end]**
[**Results_beg**] Images from a specially developed, workstation-based, 3D visualization tool are found to be an effective means of communicating the results and physics of coupled processes. The primary application of this work is hydraulic fracturing in oil or gas bearing rock. (...) The results may also be applicable to dredging, drilling and cutting of fluid saturated rock. [**Results_end]**

In this abstract, only the syntagm "Biot's theory of poroelasticity" hints at a theoretical context. The research statement does not seem to have to be justified by a gap in the existing research for instance, as presented in the CARS' model (Swales 2004).

Obviously, the pattern of moves varies across the abstracts. The four move structure presented here is nevertheless representative of a majority of the texts. Only very few do not include a research statement and a method move. Conversely and as mentioned above, contextualizing the research project, either from an institutional or a theoretical point of view, is typical of Didactics of Mathematics, while rather infrequent in Materials Science.

*Task 2: reiterated pivot keywords are identified*

For each text, the title and the research statement move were analyzed so as to identify potential keyword reiterations and the variations of their lexical environment. This was based on the assumption that the title and the research statement set the focus on a set of central concepts or keywords (Bondi 2010), which may also have been announced in the contextualization move. The title attracts the reader's attention towards a combination of keywords. The reader is then guided along the text by reiterations of these terms. The reiterated terms are included in a variation of extended lexical patterns. In most cases, the research statement move is an expansion of the title therefore repeating its central terms. It must be stated that the title used for the abstract is the title of the PhD thesis, at least when the abstract is published separately from the dissertation, which is always the case in the corpus in point.

The method used for the identification of pivot keywords included the following stages: 1) marking potential keywords in the title, 2) identifying their reiteration along the moves previously marked (see above), 3) marking these terms and their co-occurring lexical environment along the moves. Only cases where one or several keywords, first identified in the title and the research statement move, were then repeated along the moves were selected. The minimal requirement was the identification of at least one of the pivot terms in each move. The point here is not to show that all abstracts use this type of "lexical paving" based on cohesive lexical chains, but to assess their existence and their potential connection with the rhetorical dynamics. An investigation into the consequences for discursive strategy of the absence of this type of device could be the basis for further research. However, one example of a partially realized cohesive lexical chain is given and commented on in Appendix 2 (example 5 of Appendix 2).

Here is an example of this marking (in bold characters) of reiterated keywords along the rhetorical moves:

NB: only the title and the two first moves are given here.

**(3)**
[**Title_beg**] Ethnomathematics: Exploring Cultural Diversity in Mathematics [**Title_end**]
[**Research_proposal_beg**] This thesis provides a new conceptualisation of ethnomathematics which avoids some of the difficulties which emerge in the literature. In particular, work has been started on a philosophic basis for the field. [**Research_proposal_end**] [**Context_beg**] There is no consistent view of ethnomathematics in the literature. The relationship with mathematics itself has been ignored, and the philosophical and theoretical background is missing. The literature also reveals the ethnocentricity implied by ethnomathematics as a field of study based in a culture which has mathematics as a knowledge category. [**Context_end**]
[**Method_beg**] Two strategies to overcome this problem are identified: universalising the referent of 'mathematics' so that it is the same as "knowledge-making"; or using methodological techniques to minimise it. The position of ethnomathematics in relationship to anthropology, sociology, history, and politics is characterised on a matrix. (...)

The term "ethnomathematics" is first identified as the "pivot" term in the title, and then in the research statement move. It is then repeated throughout the next move within various lexical extended patterns ("conceptualization of", "view of", "position of").

All occurrences of "ethnomathematics" and its successive lexical patterns variations along the 4 moves are then listed. Only lexemes that are directly syntactically linked with the pivot term are considered (as in "position of ethnomathematics").

## 5.2. Pivot keywords: Frequency and distribution across the corpus and inside the texts

In this second step, the frequency and distribution of the terms identified across the corpora were studied using two functions of the concordancer AntConc (Anthony 2006): the "word list" and the "concordance plot". Three aspects were taken into account: 1) the frequency, inside each discipline, of the listed pivot terms; 2) their distribution across the disciplinary sub-corpus; 3) their distribution across each text.

The "wordlist" function was used to establish a frequency ranking of the identified pivot terms for each discipline. The point here was to assess how representative each pivot term was of the disciplinary ontological values: in other words, whether the listed terms pointed at central disciplinary concepts.

A general table of keywords by discipline is given in Appendix 1. Only the initial pivot terms have been listed for each text. While the 30 texts collected for each discipline do not allow for any generalisation, the list of terms can be interpreted in terms of variety and frequency. In other words, does the discipline, in the studied selection, appear as focused on a limited range of concepts or do the dissertations presented in these abstracts cover a wide variety of keywords? The data given in the table show differences between the Materials Science corpus, centered on a narrow range of keywords, such as "crack", while the Didactics of Mathematics corpus covers a much wider range of interests. The table therefore provides further grounding for the analysis of keywords as to the relationship between keywords and epistemological features (Malavasi and Mazzi 2010). The "concordance plot" function offers a graphic representation of the terms' distribution across the corpus. The pattern shows the distribution across the corpus, and the number of occurrences of the term in each text. It also gives a representation of the textual distribution of the term across each text. Here is an example for the pivot term "crack" as studied in the Materials Science corpus:

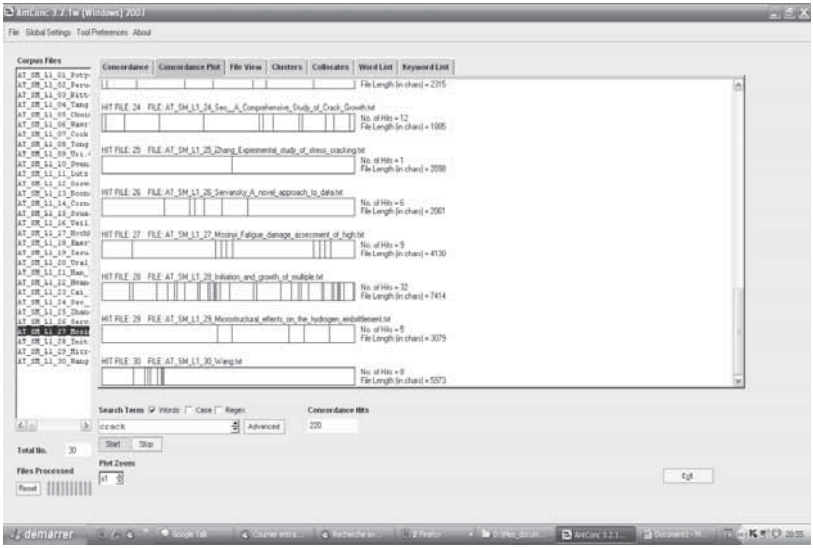



**Figure 1. "Concordance plot" for the term "crack" in the Materials Science corpus.**

Figure 1 shows that the term "crack" is widely distributed, both across the corpus of Materials Science abstracts and inside each text. "Crack" can be found in 24 of the 30 abstracts. The number of occurrences for each text range from 1 to 32, with 18 texts including more than 5 occurrences, and 5 texts including more than 10. These data tend to confirm the observation that the concept of "crack" is a core concept in the considered corpus.

### 5.3. Identification and classification of lexical variations around pivot keywords

In this third step, the variation of lexical patterns around one or several pivot terms is studied for each text, in order to ascertain the potential influence of this variation on the shift from one move to the next. This involves two successive tasks. First, each occurrence of the reiterated pivot keyword is marked in the text as well as its surrounding lexical patterns' variations (ex: "crack", "propagation", "crack propagation", "crack propagation strategy"). Secondly, the successive patterns are classified according to their distribution in the rhetorical structure (cf. Table 1).

The data thus collected at text level are compared at corpus level with a view to identifying rhetorical strategies for each discipline. The objective is to understand the extent to which individual strategies are connected with shared disciplinary values on one side, with the author's representation of the abstract as a genre, on the other side.

It must be stated here that the limited size of the corpora, while it allows identification of individual strategies at text level, implies that the described disciplinary characteristics can only be considered as indicative tendencies, which would have to be checked on a larger corpus.

## 6. Two case studies: Cohesive lexical chains or "lexical paving" in one genre and two disciplines

One example of textual collocation chain will be given here for each discipline. The 3 steps described above and their results are illustrated for each discipline. Ten more briefly commented examples are given in Appendix 2.

I will start with the Materials Science corpus.

## 6.1. A Material Science case study

- *At text level: marking of the moves and identification of the pivot terms*

The text is marked for moves. The moves 'structure of the considered abstract' is typical of its discipline (see 5.1) insofar as it only includes 3 moves, due to the absence of a contextualization move.

The title and the research statement move are studied so as to identify pivot keywords. As mentioned in 5.1, the criterion applied for a word to qualify as a pivot keyword is that it must first be identified both in the title and in the research statement. It then has to be repeated more than twice across the rhetorical moves. The example hereafter illustrates the identification of the keyword in the title and in the research statement move:

**(4)**
Title: *Computer simulation of linear and non linear* ***crack propagation*** *in cementitious materials*
Research statement: *This thesis deals with the computer simulation of* ***crack propagation*** *in cementitious materials. Both linear and nonlinear aspects of* ***crack propagation*** *are addressed.*

Here, the compound term "crack propagation" is repeated three times. "Computer simulation" and "cementitious materials" are also reiterated but only twice. They are absent in the rest of the text: therefore they have not been selected as pivot terms for the cohesive lexical chain.

- *At corpus level: frequency and distribution analysis of the terms "crack" and "propagation", using the concordancer Antconc*

The terms identified above at text level are considered at corpus level so as to ascertain their general frequency and ranking.

The "wordlist" function of Antconc is used to study the keywords "crack" and "propagation" as to their ranking in the Materials Science corpus. Here are the results for the occurrences of "crack" and "propagation" in the 30 texts:

- "crack": 220 occurrences; rank 7
- "propagation": 52 occurrences; rank 26

"Crack" is the highest ranking term of the corpus, only preceded by grammatical words, as can be seen in Figure 2.

"Propagation" is also a frequent term, ranking 26th (out of 2,540). It is the third most frequent collocate of "crack" in the corpus, after "growth" and "fatigue".

| Rank | Number of occurrences | Word |
|---|---|---|
| 1 | 936 | the |
| 2 | 551 | of |
| 3 | 374 | and |
| 4 | 339 | a |
| 5 | 286 | to |
| 6 | 267 | in |
| **7** | **220** | **crack** |

**Table 1. Frequency word list for the Materials Science corpus (11,500 tokens).**

The concordancer AntConc is then used to establish the "concordance plot" (cf. Figure 1) of the term "crack propagation" for each text. It shows that this term can be found in 6 out of 30 texts, with a wide distribution inside each text. The word "crack" can be found in 24 texts, "propagation" in 16 texts. The compound term "crack propagation" and its two components can therefore be considered as widely distributed.

- *Back to the text level: identification and classification of lexical variations around pivot keywords*

| Keywords | Crack Propagation |
|---|---|
| Title | Computer simulation of linear and non linear **crack propagation** in cementitious materials |
| Research statement | Crack propagation (2 occ.)<br>Crack propagation process (2 occ.) |
| Method | Crack propagation<br>Crack model<br>Crack propagation strategy<br>Criterion for propagation<br>Directionof propagation<br>Propagation length<br>Cohesive crack problem |
| Results | Crack propagation process |

**Table 2. Lexical patterns variations around pivot keywords in a Materials Science PhD abstract.**

The two identified keywords ("crack" and "propagation") and their pattern's variations throughout the abstract are identified and classified as to their distribution across the rhetorical structure.

The initial lexical pattern "crack propagation" is first repeated then expanded ("crack propagation process") in the research statement move. As mentioned above, there is no contextualization move in this text. In the method move, the same reiteration and expansion can be found; however, the two initial pivot terms ("crack" and "propagation") are successively associated and dissociated, with the keyword "crack" opening and closing the variation. The results move repeats the expanded form ("crack propagation process") found in the initial proposition move. The global picture is one of a sort of folding and unfolding process, or "packing" and "unpacking" (Halliday 1998) along the moves. Here, this process is based on the compounding and un-compounding of a general scientific term ("process", "strategy", "criterion"…) with a field-specific term ("crack", "propagation").

The lexical pattern's variation folds and unfolds as if reproducing the heuristic process which leads from focusing on a given (folded) phenomenon (here "crack propagation process") to its explanation (unfolding) through experimentation and finally to the closure of both the text and the experimental process with the reiteration of the initial pattern (folding). This type of cohesive lexical chain is based on the reiteration of two pivot keywords that appear as representative of a disciplinary focus of interest, at least as seen from this micro-corpus.

## 6.2. A Didactics of Mathematics case study

Let us move on to consider the Didactics of Mathematics corpus. As for the previous case study in Materials Science, one of the 30 abstracts taken from Didactics of Mathematics is studied here in order to locate and then ascertain the role of a cohesive lexical chain or "lexical paving". The same approach is used to 1) identify rhetorical moves and pivot keywords 2) assess the frequency and distribution of these pivot keywords 3) analyse the interaction between the keywords' combination patterns and the rhetorical structure at text level.

- *At text level: identification of the pivot terms, starting from the title and the research statement*

The text is marked for moves. The title and the four moves are analysed as to identify a lexical reiteration. Reiterated terms are identified

as pivot keywords, starting from the title and the research statement. As mentioned in 5.1, the criterion applied for a word to qualify as a pivot keyword is that it must first be identified both in the title and in the research statement. It then has to be repeated more than twice across the rhetorical moves. The example hereafter illustrates the identification of the keyword in the title and in the research statement move:

**(5)**
Title: ***Ethnomathematics****: Exploring Cultural Diversity in Mathematics*
Research statement: *This thesis provides a new conceptualisation of* ***ethnomathematics*** *which avoids some of the difficulties which emerge in the literature*

Here, the term "ethnomathematics" appears alone in the title; it is then introduced by "a new conceptualisation of". "Ethnomathematics" is reiterated several times along the moves, with various lexical environments. It is therefore considered as the pivot term.

- *At corpus level: frequency and distribution analysis of the term "ethnomathematics"*

The identified pivot keyword frequency is studied using the AntConc wordlist function, which shows a ranking of 165 for "ethnomathematics" with only 11 occurrences of this term. As compared to "crack"' in the Materials Science corpus studied above, "ethnomathematics" has a low ranking.

The "concordance plot" function (cf. Figure 1) is then used to study the distribution of the term across the texts of this corpus: "ethnomathematics" appears only in one abstract where it can be found 11 times. As seen from this micro-corpus, it can therefore be considered as representative of a specific research project, but not as a core concept of the discipline.

- *Back to the text level: lexical pattern variations along the rhetorical move structure*

The various lexical extended forms around the pivot keyword "ethnomathematics" are identified and classified according to their distribution across the rhetorical move structure, as shown in Table 3.

| Keyword | Ethnomathematics |
|---|---|
| Title | Ethnomathematics: Exploring Cultural Diversity in Mathematics |
| Research statement | Conceptualisation of ethnomathematics |
| Context | View of ethnomathematics, ethnomathematics |
| Method | Position for ethnomathematics<br>Place for ethnomathematics<br>Ethnomathematical activity<br>Ethnomathematical work<br>Ethnomathematical theory |
| Results | Ethnomathematics (3 occ.) |

**Table 3. Lexical patterns variations around pivot keywords in a Didactics of Mathematics PhD abstract.**

The term "ethnomathematics" is announced in the title then used again in the research statement move, with the adjunction of the general scientific term "conceptualization". The same pattern can be found in the contextualization move with the adjunction of the term "view". In the method move, two patterns are adopted: the first one combines a general term ("position", "place") and the pivot term; the second pattern transforms the grammatical status of the pivot term, changing it into the adjective "ethnomathematical". This adjective modifies three general terms "activity", "work", "theory". The abstract finally closes with the simple term "ethnomathematics" in the results move.

The global picture is similar to the one in the Materials Science corpus as the lexical variation follows a lexical adjunction and lexical disjunction process. Just as in the Materials Science corpus, the lexical variation seems to evolve along the rhetorical structure, alternatively expanding and condensing the lexical pattern along the moves

However, this pattern is not based here on compounding and uncompounding, but rather on the adjunction of a prepositional segment including a general term such as "view of", "position for", "place for". Another difference from the Materials Science case is that, based on the frequency and distribution analysis, the pivot keyword "ethnomathematics" appears as pointing to the focus of specific research rather than a core disciplinary concept.

## 7. One generic pattern and two disciplinary rhetoric strategies

The vast majority of cohesive lexical chains or "lexical paving" identified in the two disciplinary corpora are based on the combination of general scientific terms (e.g. "conception", "model", "approach") and domain-specific terms as in the two case studies above. However, the lexical profile appears quite different for each discipline. The lexical chains in the Materials Science corpus are organised around a narrow range of pivot terms such as "crack" and "propagation". Conversely, in the Didactics of Mathematics corpus, the pivot terms are extremely varied, with most of them specific to one text. The only recurrent pivot term across the texts of the corpus is "mathematics" which could be expected, considering the domain. In Materials Science, only a few general scientific terms are used as pivot terms with one very recurrent term ("simulation" in 6 texts out of 15); their range is much wider in Didactics of Mathematics, with only one general term ("achievement") repeated in 2 lexical chains.

In the two disciplines, and in the majority of the abstracts, the lexical pattern variations around reiterated pivot terms follow a sort of "folding and unfolding" process starting from the title and evolving along the rhetorical moves. In a first step, the focus is set on one or two combined pivot terms, often specialised ones (e.g. "ethnomathematics", "crack", "propagation"). These terms are then repeated while being included in a varying lexical environment, in an expansion process (e.g. "ethnomathematics followed by "view of ethnomathematics", "position of ethnomathematics"; "crack propagation" followed by "crack propagation process", "crack propagation strategy"). The results move closes the text on a final condensation, returning to or combining the initial pivot terms (see lexical items in bold characters in Appendix 2). As mentioned by Pecman (2012) in a study of emerging neologisms in scientific papers, this type of discursive pattern can be accounted for using the Hallidayan approach of information structure, as based on a distinction between "given" and "new" facts (Halliday and Hasan 1976). In this case, the first "folded" occurrence of the term should be considered as the "given" fact, while its "unfolded" (or expanded) form would show the concept under a new light, so as to explain it. This type of variation may also be considered as an illustration of the use of grammatical metaphors to transform an assumption into accepted knowledge (Halliday 2004). From that point of view, the initial presentation within a single lexical item of one or several combined pivot keywords, would aim at presenting the object of research

as a given fact. The following unfolding process, through lexical patterns' variations, would show various aspects of the studied concepts, considered as new. The final condensation or "folding" would therefore mark the end of the demonstration and its conclusion, "creating the effect of 'emerging knowledge', which ultimately draws the reader into the world of scientific discovery" (Pecman 2012, 42). The described process of lexical variation has implications as to the understanding of the term formation process as well as to specific discursive strategies (Humbley 2009).

However, this "folding" and "unfolding" pattern is not based on the same type of lexical variation in the two disciplines. In Materials Science, a relatively narrow range of specialised terms (Appendix 1) tends to be combined with an even narrower range of general scientific terms in varying and complex compound expressions (ex: "arbitrary cohesive crack propagation strategy"). Didactics of Mathematics shows a preference for a wide range of single specialised terms generally included in a prepositional pattern. The variation process is mostly based on a sequence of adjuncts such as "conceptualization of", "view of", "position of". Considering the specific objectives of each discipline, it may be assumed that Materials Science describes and explains a phenomenon by simulating and observing it: the lexical pattern variation "mimics" the folding and unfolding hermeneutic process which discloses the mechanism of the observed phenomenon. On the other hand, Didactics of Mathematics aims at transforming an initial situation to improve it: the focus is initially set on a didactic objective or method. It is then seen from various angles, considering possible transformations. The lexical variation in the cohesive chain is based on adjunctive patterns which express the evolving point of view. This difference in the type of lexical pattern variations is coherent with the differences in the moves' distribution between the two disciplines. The very short space devoted to theoretical contextualization in Materials Science can be related to an objectivist experimental pattern where science is built on successive experimentation (Flöttum, Dahl and Kinn 2006). By contrast, writers in the Didactics of Mathematics use up to a third of their abstracts to present the institutional and theoretical context of their research. The argumentation that follows, based on an interpretative model, aims to transform the initial didactic context through the application of a new concept.

Therefore it appears that PhD abstracts share a common rhetorical and lexical pattern, which involves the construction of a cohesive lexical chain, based on a succession of lexical pattern variations around reiterated pivot keywords or "lexical paving". Despite individual specificities, the prevalent variation pattern evolves from condensation to expansion and

back to condensation, as the focus shifts from contextualization and research statement to method and results. The disciplinary specificities are visible in the lexical choice, the balance between general and specialised pivot terms, and the type of lexical variation patterns around these terms.

## 8. Conclusion

The study of lexical variations along the rhetorical structure of PhD abstracts in the fields of Didactics of Mathematics and Materials Science provides evidence that the identified cohesive lexical chains contribute to the creation of a textual dynamic. The reiterated keywords and the variation of their lexical environment guide the reader's attention along the moves, starting from a research statement and ending with results and perspectives, in a sort of folding and unfolding process. Insofar as it ensures an efficient cohesive effect, this rhetorical device contributes to the credibility of the research project and of its author, which is one of the main objectives of this "self-promotional tool" (Hyland 2004). However, the selected type of lexical variation seems to be connected with the epistemological values that are specific to the discipline. While the folding and unfolding pattern appears as characteristic of most cohesive lexical chains in abstracts (cf. Appendix 2), the type of lexical combination within this pattern obeys a disciplinary model which may be experimental or argumentative, as is the case of Materials Science and Didactics of Mathematics The choice of lexical patterns appropriate as to the discipline's epistemology may be considered as a contribution to the writer's legitimacy as a "would-be insider" (Hyland 2004) of a specialised discourse community.

Therefore, while disciplinary variations might seem to make the limits of the genre more fuzzy, they actually make the PhD abstract more efficient in reaching the genre's target i.e. convincing the specific disciplinary community that the abstract's authoris a worthy candidate as a future researcher. As Swales (1990, 49) puts it, "exemplars or instances of genres vary in their prototypicality". The comparative study of such text-structuring lexicogrammatical features as "lexical paving" in various disciplines helps us to understand how writers remain in conformity with the norms of the genre while "appealing to readers from within the boundaries of a disciplinary discourse" (Hyland 2004, 63). It also opens the way to further comparative research as to the use of this cohesive device in various genres, such as review and research papers.

## Notes

[1] Acknowledgements to Mojca Pecman for this suggestion.
[2] For example, Elsevier's Scopus http://www. info.sciverse.com/scopus
[3] http://www.ndltd.org/serviceproviders/scirus-etd-search

## References

Anthony, Laurence. 2006. Developing a freeware, multiplatform corpus analysis toolkit for the technical writing classroom. *IEEE Transactions on Professional Communication* 49(3): 275-286.

Bhatia, Vijay K. 1993. *Analysing genre: Language use in professional settings*. London: Longman

Biber, Douglas, Ulla Connor, and Thomas Upton. 2007. *Discourse on the move*. Amsterdam: John Benjamins.

Bondi, Marina. 2002. Attitude and episteme in academic discourse: Adverbials of stance across genres and moves. *Textus* 15(2): 249-264.

—. 2010. Perspectives on keywords and keyness. In *Keyness in texts*, ed. Marina Bondi and Mike Scott, 5-19. Amsterdam: John Benjamins.

Biehler, Rolf, Roland W. Scholz, Rudolf Sträßer, and Bernard Winkelmann. ed. 1994. *Didactics of mathematics as a scientific discipline*. Dordrecht: Kluwer.

Bordet, Geneviève. 2014. Influence of collocational variations on making the PhD abstract an effective "would-be insider" self-promotional tool. In *Abstracts in academic discourse: Variation and change*, ed. Marina Bondi and Rosa Lorés Sanz, 131-160. Bern: Peter Lang.

Charles, Maggie. 2003. "This mystery...": A corpus-based study of the use of nouns to construct stance in theses from two contrasting disciplines. *Journal of English for Academic Purposes* 2(4): 313-326.

Dressen-Hammouda, Dacia. 2008. From novice to disciplinary expert: Disciplinary identity and genre mastery. *English for Specific Purposes* 27(2): 233-252.

Firth, John Rupert. 1957. *Papers in linguistics, 1934-1951*. Oxford: Oxford University Press.

Fløttum, Kjersti, Trine Dahl, and Torodd Kinn. 2006. *Academic voices across languages and disciplines*. Amsterdam: John Benjamins.

Gledhill, Christopher. 1995. Collocation and genre analysis. The phraseology of grammatical items in cancer research articles and abstracts. *Zeitschrift für Anglistik und Amerikanistik* XLIII 1(1): 11-36.

—. 2000. The discourse function of collocation in research article introductions. *English for Specific Purposes* 19(2): 115-135.
—. 2011. The 'lexicogrammar' approach to analysing phraseology and collocation in ESP texts. *ASp. La revue du GERAS* 59: 5-23.
Halliday, Michael. A. K. 1998. Language and knowledge: The 'unpacking'of text. In *Text in education and society*, ed. Allison Desmond, Wee Lionel, Zhiming Bao, and Anne Abraham Sunita Anne, 157-178. Singapore: Singapore University Press.
—. 2004. *The language of science*. London: Continuum.
Halliday, Michael. A. K., and Ruqaiya, Hasan. 1976. *Cohesion in English*. Harlow: Longman.
Hoey, Michael. 2005. *Lexical priming*. London: Routledge.
Huckin, Thomas. 2001. Abstracting from abstracts. In *Academic writing in context: Implications and applications*, ed. Martin Hewings, 93-103. Birmingham: Birmingham University Press.
Humbley, John. 2009. Accounting for term formation. *Terminology Science and Research* 20: 1-15.
Hyland, Ken. 2000. *Disciplinary discourses: Social interactions in academic writing*. London: Longman.
—. 2002. Genre: Language, context, and literacy. *Annual Review of Applied Linguistics* 22: 113-135.
Malavasi, Donatella, and Davide Mazzi. 2010. History v. marketing: Keywords as a clue to disciplinary epistemology. In *Keyness in texts*, ed. Marina Bondi and Mike Scott, 169-184. Amsterdam: John Benjamins.
Partington, Alan. 1998. *Patterns and meanings: Using corpora for English language research and teaching*. Amsterdam: John Benjamins.
Pecman, Mojca. 2012. Tentativeness in term formation: A study of neology as a rhetorical device in scientific papers. *Terminology* 18(1): 27-58.
Renouf, Antoinette, and John McH. Sinclair. 1991. Collocational frameworks in English. In *English corpus linguistics: Studies in honour of Jan Svartvik*, ed. Karin Aijmer and Bengt Altenberg, 128-143. London: Longman.
Swales, John. 1990. *Genre analysis: English in academic and research settings*. Cambridge: Cambridge University Press.
Sinclair, John McH. 1991. *Corpus, concordance, collocation*. Oxford: Oxford University Press.

Swales, John. 2004. *Research genres: Explorations and applications*. Cambridge: CUP.

Swales, John, and Christine Feak. 2009. *Abstracts and the writing of abstracts*. The Michigan Series in English for Academic and Professional Purposes. Ann Arbor: University of Michigan Press.

Teubert, Wolfgang. 2007. Corpus linguistics and lexicography. *Text Corpora and Multilingual Lexicography* 8: 109-133.

## Appendix 1: Pivotal keywords table

Below, for each discipline, a list of reiterated pivotal keywords, in alphabetical order. Terms between brackets are recurrent collocates of the pivot term in the abstract.

| Materials Science | Didactics of mathematics |
|---|---|
| boundary element analysis | ability tracking |
| (elastic stress) waves | achievement |
| (fatigue) crack | agricultural education |
| (fatigue) crack | Appalachia |
| (stress) crack | Assessment |
| asphalt concrete | assessment (renewal) |
| cohesive | calculators and computers |
| concrete | centralist |
| crack | class discussion |
| crack | Constructivist |
| crack (growth) | diagnostic inventory |
| (discrete) crack propagation | discrete mathematics |
| crack (propagation) | (previous mathematics) experiences |
| crack growth | Ethnomathematics |
| crack growth | Fractions |
| crack propagation | Game |
| crack tip | Gender |
| damage | Integration |
| (multiple-site) damage | life histories |
| delamination | logarithms and logarithmic functions |
| electric current | manipulative use |
| fatigue | math talk |
| fatigue crack | mathematical knowledge |
| fatigue crack | Mathematics |
| fatigue crack (growth) | Mathematics |
| grain boundary | mathematics (achievement) |
| hydraulic fracturing | mathematics classes |
| hydrogen | mathematics curriculum materials |

| | |
|---|---|
| interface | mathematics integration |
| microstructure | mental (computation) |
| microstructure | Multiple intelligences (learning) |
| model | (teaching and learning) package |
| numerical | preservice teacher |
| oxidation | preservice teachers |
| pipe | preservice teachers |
| poroelastic | program components |
| residual strength | proof (scheme) |
| simulating | prospective teachers |
| two and three dimension | Reform |
| slabs | ritual |
| stress | self-efficacy |
| | student (mathematics) learning |
| | student outcomes |
| | teacher education |
| | teacher variables |
| | teachers' learning |
| | teachers' knowledge |
| | technical students |
| | technological problem-solving |
| | Technology Education, Science, and Mathematics Integration |

# Appendix 2: Case studies

The examples below have been studied so as to identify the reiteration pattern of pivot keywords inside 10 texts, taken respectively from the didactics of mathematics and materials science corpora.

For each text, a table lists:

- the pivot keywords
- the title
- the occurrences of the pivot keywords in each of the 4 moves

For each text, the interaction between the moves' structure and the pivot keywords reiteration's pattern is commented as to the "folding" and "unfolding" process presented in the study. Examples 4 and 5 present variants as to this type of discursive pattern.

The lexical items marked in bold characters offer a combination of all the pivot keywords, as an illustration of the "folding" process.The pivot keywords are underlined in the title.

**Examples taken from the didactics of mathematics corpus**

**(1)**

| | |
|---|---|
| **Pivot keywords** | - Preservice teachers |
| | - mathematics curriculum material |
| **Title** | **<u>Preservice</u>Elementary <u>Teachers</u>' Learning with <u>Mathematics Curriculum Materials</u> During Preservice Teacher Education** |
| **Context** | - development of curriculum materials |
| | - "Standards-based" curriculum materials |
| | - teachers' experiences with these materials |
| | - **preservice teachers' use of mathematics curriculum materials** |
| **Research Proposal** | - **preservice teachers' interactions with mathematics curriculum materials** |
| **Method** | - preservice elementary teachers |
| | - mathematics curriculum materials and textbooks |
| | - mathematics curriculum materials and textbooks |
| | - **preservice teachers' initial interpretations of Standards-based curriculum materials** |
| | - preservice teachers' experiences |
| | - mathematics curriculum materials |
| | - preservice teacher learning |
| | - Standards-based curriculum materials |

| | |
|---|---|
| **Results** | - preservice teachers |
| | - mathematics curriculum materials |
| | - many of the materials |
| | - preservice teachers |
| | - materials |
| | - curriculum materials |
| | - preservice teachers |
| | - **preservice teachers' encounters with mathematics curriculum materials** |

Comment: the text is based on the interrelation of pivot terms pointing to, respectively, education actors ("preservice teachers") and educational resources ("mathematics curriculum material"). The title and each move include an extended lexical item combining both pivot terms.

**(2)**

| | |
|---|---|
| **Pivot Keywords** | - teacher |
| | - program |
| **Title** | **Preservice Teachers' Characterizations of the Relationships Between Teacher Education Program Components** |
| **Research Proposal** | - **elementary teacher education program** |
| | - preservice teachers |
| | - preservice teachers |
| | - **teacher education program components** |
| | - **program components to teaching** |
| **Method** | - preservice teachers |
| | - program emphases |
| | - program recommendations |
| | - program-based philosophies |
| | - preservice teachers |
| | - program components |
| | - the accomplishment of program objectives |
| | - program objectives proved unrealistic |
| **Results** | - program intentions |
| | - program suggestions |
| | - program-recommended practices |
| | - preservice teachers' learning |
| | - **teacher education program coursework** |

Comment: the text is based on the interrelation of pivot terms pointing to education actors ("teachers") and educational resources ("program"). The text opens and closes with a complex lexical item including both elements.

**(3)**

| | |
|---|---|
| **Pivot keywords** | - teacher variables |
| | - student mathematics learning |
| | - manipulative use |
| **Title** | **Teacher Variables and Student Mathematics Learning Related to Manipulative Use** |
| **Context** | - teacher background variables |
| | - student learning |
| | - these variables |
| | - teacher variables |
| | - student learning |
| | - use of manipulatives |
| **Research Proposal** | - **role of manipulative use as a mediator of the relationship between teacher variables and students' mathematics learning** |
| **Method** | - manipulative use |
| | - teacher variables |
| | - manipulatives |
| | - manipulative use |
| | - manipulatives |
| | - teacher variables |
| | - manipulative use |
| | - manipulative use |
| | - mathematics learning |
| **Results** | - manipulative use |
| | - student mathematics learning |
| | - teacher variables |
| | - student learning |
| | - manipulatives |
| | - manipulatives |
| | - manipulatives |
| | - mathematical learning |
| | - **relationship between teacher variables and student learning** |
| | - teacher variables |
| | - manipulative use |
| | - students' mathematics learning |
| | - manipulative use |
| | - **manipulative use in the teaching and learning of mathematics** |

Comment: the text is based on the interrelation between 3 pivot terms pointing to 3 concepts: "teachers' variables", "students' mathematics learning" and "manipulative use". The 3 concepts are interconnected in the

title and in the research proposal move. They are only connected by pairs in the results move.

**(4)**

| | |
|---|---|
| **Pivot keywords** | - constructivist |
| | - centralist |
| **Title** | An interpretive study of the role of teacher beliefs in the implementation of **constructivist** theory in a secondary school mathematics classroom |
| **Research Proposal** | - A constructivist-related theory |
| | - teachers' centralist classroom roles |
| **Method** | - a constructivist-oriented teaching approach |
| | - teacher's centralist pedagogy |
| | - underpinning constructivist theory |
| | - centralist classroom role of teacher as informer |
| | - multiple constructivist-related perspectives (radical constructivism, social constructivism, critical constructivism) |
| **Results** | - the teacher's refined centralist classroom role of teacher |
| | - centralist classroom role of teacher as controller |
| | - cognitivist theory of constructivism |
| | - future constructivist-related pedagogical reform |
| | - a critical constructivist perspective |

Comment: only the term "constructivist" is present in the title. Starting from the research proposal move, the text is based on the duality opposing "centralism" and "constructivism". The pivot keyword "centralist" combines with "teacher" and "classroom" while "constructivist" combines with general terms such as "theory" and approach". "Constructivism" opens and closes the text, thus setting the main focus, based on only one pivot keyword. The pivot term "centralist", although it can be found in each move, is not present in the title and is never combined with the pivot term "constructivist. Therefore it appears as a mere counterpoint of the main focus, set on "constructivism".

**(5)**

| | |
|---|---|
| **Pivot keywords** | - discussion<br>- reform |
| **Title** | **Class <u>Discussion</u>: One Teacher's Struggles in Implementing <u>Reform</u>** |
| **Research Proposal** | - mathematical discussions |
| **Method** | - class discussion |
| | - patterns of discussion |
| | - discussion |
| | - reform-oriented strategies |
| | - **reform-oriented strategies for discussion** |
| | - teacher's vision of reform |
| | - for reforming |
| | - whole-class discussions |
| **Results** | - **to reform discussion in his classroom** |
| | - class discussions |
| | - current reform initiatives |

Comment: this example is atypical: the two pivot terms pointing at central concepts ("class discussion" and "reform") are first announced in the title then only combined in a unique lexical pattern in the method move and in the results move but not in the research proposal. There is no contextualization. This textual pattern seems to weaken the dynamics, and the persuasive potential, since the reader's attention is not immediately focalized on the actual research issue, which is not either justified by a contextualization

**Examples taken from the materials of science corpus**

**(6)**

| | |
|---|---|
| **Pivot keywords** | Simulate |
| | Crack |
| **Title** | **A software framework for simulating curvilinear crack growth in pressurized thin shells** |
| **Research Proposal** | **simulating crack growth** |
| **Method** | crack trajectories |
| | Simulation |
| | a fracture simulation code |
| | the fracture simulation code |
| | simulation attributes |
| | crack growth |
| | Cracks |
| | crack growth results |
| | crack growth |
| | modified crack closure integral method. |
| | crack trajectory |
| **Results** | **simulating crack growth** |
| | fracture simulation code |
| | crack trajectory |
| | total crack length |
| | predicted crack trajectory |

Comment: the interaction between "simulation" and "crack" is the foundation of the text. The two lexical items are combined in a unique lexical pattern, both in the research proposal and in the results move.

**(7)**

| | |
|---|---|
| **Pivot keywords** | - discrete |
| | - crack |
| | - two and three dimensions |
| **Title** | **Discrete modeling of crack propagation: theoretical aspects and implementation issues in two and three dimensions** |
| **Research Proposal** | - **discrete crack propagation for two- and three-dimensional problems** |
| **Context** | - discrete modeling of crack propagation |
| | - individual cracks |
| **Research Proposal** | - arbitrary crack growth |
| **Method** | - the evaluation of crack-tip parameters, crack stability, crack propagation, initial stresses, interfacial cracks |
| | - two- and three-dimensional crack propagation analyses |
| **Results** | - three-dimensional program |
| | - **three-dimensional discrete crack propagation simulations** |

Comment: the three terms ("discrete", "crack", "three dimensional") are used in an extended lexical item in the research proposal: the concepts are then dissociated in the method move and reconnected in a single lexical pattern at the end of the text.

**(8)**

| | |
|---|---|
| **Pivot keywords** | - surface flaw |
| | - cyclic |
| **Title** | **Three-dimensional finite element analysis of cyclic fatigue crack growth of multiple surface flaws** |
| **Research Proposal** | - **multiple surface flaws in a plate subjected to cyclic loading** |
| **Method** | - flaw interaction effects |
| | - **surface flaws subjected to cyclic loading** |
| | - single surface flaw |
| | - cyclic stress-intensity factors |
| | - cyclic crack growth properties |
| | - **interacting surface flaws in a plate subjected to cyclic loading** |
| **Results** | - **multiple surface flaws in a plate subjected to cyclic loading** |
| | - a single surface flaw |
| | - **interacting surface flaws subjected to cyclic loading** |

Comment: the two pivot terms ("cyclic", "surface flaw") are reiterated both separately and as combined in a unique extended lexical item, for each move.

**(9)**

| | |
|---|---|
| **Pivot keywords** | - crack |
| | - numerical |
| **Title** | **Virtual crack extension method for calculating rates of energy release rateand numerical simulation of crack growth in two and three dimensions** |
| **Research Proposition** | - virtual crack extension method |
| | - **a numerical procedure for simulating a growth of multiple crack systems** |
| **Method** | - crack extension method |
| | - crack extension for multiply cracked bodies |
| | - multiple crack systems |
| | - crack-face |
| | - multiple crack systems |
| | - crack-growth model |
| | - numerical procedure |
| | - planar cracks |
| | - virtual crack extension method |
| | - crack |
| **Results** | - propagating crack |
| | - crack extensions |
| | - crack front |
| | - crack front |
| | - virtual crack extension method |
| **Method** | - approximate numerical procedure |
| | - non-straight cracks |
| **Results** | - **numerical simulation of inclined central cracks** |
| | - crack surface |
| | - reasonable crack-growth pattern |

Comment: the two keywords ("numerical" and "crack") are combined inside one lexical pattern both in the opening research proposal move and the closing results move. It must be noticed that the term "virtual" is present in the title, the research proposal, the first "method" step and the first "result" step but not in the final part of the method and the results move. The abstract actually follows a two-step pattern, with one based on the interaction between "virtual" and crack", and the other on the interaction between "numerical" and "crack".

**(10)**

| | |
|---|---|
| **Pivot keywords** | - growth |
| | - damage |
| **Title** | **Initiation and growth of multiple-site damage in the riveted lap joint of a curved stiffened fuselage panel: an experimental and analytical study** |
| **Research Proposal** | - **multiple-site damage (MSD) initiation and growth** |
| **Method** | - crack growth history |
| | - crack formation and growth |
| **Results** | - MSD cracks |
| **Method** | - crack initiation and growth |
| | - damage and crack initiation |
| **Results** | - crack growth |
| **Method** | - fatigue crack growth |
| | - NASGRO crack growth model |
| | - crack growth rate |
| **Results** | - fatigue crack growth predictions |
| | - crack growth data |
| | - **growth of MSD cracks** |
| | - **MSD crack initiation and growth** |

Comment: the two pivot terms ("damage" and "growth") are combined in a single lexical pattern in the opening research proposal move and in the closing results move. "Damage" is first presented within an expanded acronym (multiple-site damage), then inside the acronym (MSD).

# CHAPTER THREE

# RESEARCH ARTICLES IN SOCIOLOGY: VARIATION WITHIN THE DISCIPLINE

ŠAROLTA GODNIČ VIČIČ
UNIVERSITY OF PRIMORSKA
AND MOJCA JARC
UNIVERSITY OF LJUBLJANA, SLOVENIA

## 1. Introduction

The ways in which academic communities disseminate knowledge through research articles (RAs) has been extensively studied (for an overview see Hyland 2006; Hyland and Salager-Meyer 2008). The majority of genre-based studies of RAs focus on variation in linguistic phenomena from a cross-disciplinary perspective. Disciplines are mostly compared across the soft vs. the hard sciences divide (e.g. Hyland 2001; Charles 2007). Sometimes related disciplines are compared; for example, political science, sociology and history (Holmes 1997), or conservation biology and wildlife behaviour (Samraj 2005). Studies also focus on individual disciplines – e.g. linguistics (Ruiying and Allison 2003; Lorés Sanz 2004), medicine (Luzón Marco 2000), pharmaceutical sciences (Gledhill 2000), economics (Dahl 2009), history (Bondi 2009a), computer sciences (Posteguillo 1999) and biochemistry (Kanoksilapathan 2005) – aiming at determining the rhetorical structure of RAs in the discipline or the typical phraseology used in them. Some of the studies of RAs of a single discipline take a contrastive approach and focus on cross-cultural variation (Salager-Meyer et al. 2003; Bondi 2009b; Mur Dueñas 2010; Lorés Sanz 2011).

In cross-disciplinary studies, variability in RAs tends to be attributed to variation in disciplinary norms, differing communication needs and communicative purposes of the academic communities in question. The

explanation of variability within a single discipline seems less straightforward. Sometimes it is not commented on, probably as it is assumed that variability within the discipline is due to the stylistic flexibility with which authors use language to transform experience or observation into knowledge. Flexibility in the structure of moves pertaining to the genre of RA sections is sometimes explained by the author's position in the academic community or by the different communicative functions that RAs can have within a discipline (Lorés Sanz 2004).

Ozturk (2009) has addressed variation within a single discipline in a non-cross-cultural manner. He compared the generic moves in the introduction sections of RAs in two subdisciplines of applied linguistics; i.e. second language acquisition and second language writing research, following Swales' (1990) CARS model. Variation in the way RA introductions are organised in the two subdisciplines was attributed to their different maturities: the former being a well-established subdiscipline, the latter an emerging one. Ozturk's pioneering research, however, was small-scale and the extent to which his findings may be applied to other fields and their subdisciplines remains an open question.

In an attempt to provide further insight into intradisciplinary variation in RAs, this study will focus on a well-established discipline: sociology. Taking an integrative approach which aims to link linguistic evidence of variation in RAs with the context in which these RAs are created, published or consumed, the chapter begins with a brief overview of recent literature on disciplinary characteristics of sociology and sociologists' publishing practices as well as on the RA genre in the field of sociology (section 2). This is followed by a description of the corpus and methodology used in the current study (section 3). Sections 4 and 5 discuss the findings emerged from the analysis.

## 2. Sociology, sociologists and scholarly writing

Sociology as a social science is systematically engaged in studying human society and is concerned with questions such as social order, conflict and change, power, inequality and social reproduction, and so on. Increased fragmentation and specialization around research topics have characterized sociology throughout its history, causing its boundaries to change and converge in new hybrid fields of inquiry. Moreover, the differentiation of the discipline has been coupled with pluralism of theoretical and methodological orientations. Although there were periods when certain paradigms prevailed, a central paradigm has not been

established and sociological endeavours have been nourished by various concepts, methods and models. The lack of a common paradigmatic and methodological core has been felt by sociologists as both an advantage and one of the basic tensions of their work (Quah and Sales 2000). Dogan (2000) maintains that sociology remains one of the most open of disciplines. Communication across disciplinary borders with such fields as anthropology, political science, economics, philosophy, psychology, and even mathematics and physics has been common from early on. Furthermore, sociologists often migrate both to and from these disciplines, exchanging concepts, theories and methods. Pontille (2003) finds that migrating scholars with initial education in another discipline are often those who introduce change in sociologists' methods of research and into their writing practices. The objects of sociological inquiry, nonetheless, tend to be more affected by institutionalization processes and funding opportunities available for research.

Sociologists tend to operate in a number of different academic communities at the same time; their perceptions of these communities may nevertheless differ across national contexts and depend not only on individual institutional, national and regional contexts (Pontille 2003), but also on the research topics they investigate (Jarc and Godnič Vičič 2012). Vanderstraeten (2010) suggests that greater access to scientific communication networks and publications has opened up space for collaboration around research topics and introduced change in publication practices in sociology. As a result, the average number of co-authored articles as well as the number of female authors has increased.

Publication practices of sociologists largely depend on their local institutional and social contexts. While only a decade ago Quah and Sales (2000) suggested that sociologists living outside Northern America and Western Europe might have limited access to the publications of their western peers and their own research was often kept within national borders, the number of sociology journals from non-English speaking countries included in the *Social Sciences Citation Index* (SSCI) has recently increased (Testa 2011). Sociologists from countries where English is a foreign language increasingly use English as the *lingua franca* of their discipline: they regularly use it for reading and writing RAs, participating at international conferences and working in international project teams. They publish RAs not only in their own language, but also in other languages. Their research articles in English are published in national English-medium journals, in English-medium journals with a regional scope, and in Anglophone journals with a global scope. The decision about which journal a RA should be sent to does not depend only

on the research topic and methods of the research used but also closely relates to the niches the journals occupy, the recommendations of peers, as well as the opportunities arising from collaborations with others (Jarc and Godnič Vičič 2012).

RAs seem to reflect the fragmented nature of sociology in different ways. Pontille (2003) maintains that the materials and methods used in a study affect the rhetorical structure of the RA, and Harwood (2009) believes that the ways in which sociologists use citations depends on text type.

The linguistic aspects of sociology RAs have mostly been dealt with from a cross-disciplinary perspective, the only notable exception being Brett (1994), who found that introductions of RAs are not always titled, with the result that these sections often contain more procedural information and substantiation than RAs of other disciplines. In a cross-disciplinary tradition, Holmes (1997) compared sociology RAs with other soft sciences and Bruce (2009) with organic chemistry. While Holmes found that discussion sections in sociology and political science are similar to those in the hard sciences if less complex and less predictable as regards the presence and order of generic moves, Bruce noticed that the rhetorical structure of results sections varies across RAs, and that authors often use amplification and headings to subsections that report individual findings. He also noticed that the style of writing in sociology can be personal or impersonal but has to allow for multiple meanings and views. Sociology RAs have been compared to both soft and hard sciences by Hyland in a number of studies (e.g. 1999, 2007, 2008). He found that sociologists employ the highest number of citations in their RAs as opposed to other disciplines; however they show little variation in the ways they employ them. Like Brett, he, too, noticed that exemplifying plays an important role in sociologists' attempts to contextualize propositions and engage with readers. Interaction with readers is also extensively realised by stance and engagement markers which participate in the construction of authors' identities and reader persuasion regarding the legitimacy of authors' knowledge claims.

## 3. Methods and materials

It seems that sociologists take variation within their discipline for granted while linguists (with the exception of Holmes 1997 and Harwood 2009) have scarcely noticed the phenomenon. To address this lack of linguistic research, this corpus-based study aims to establish whether the

language of RAs reflects the intra-disciplinary variation in the field of sociology.

Due to the multiplicity of paradigms, theoretical and methodological approaches and specializations within the field of sociology, the migrations of scholars to sociology from other disciplines, as well as institutional, national and international circumstances in which sociological knowledge is created and disseminated, it seems that too many factors are at play to control the variables that can affect variation in RAs within this field of inquiry. Thus, a corpus consisting of a random sample of RAs from several sociology journals, which is often used as a corpus compilation method in linguistic studies, would probably not reflect the diversity of sociologists' lexical primings (Hoey 2004) and as such it would not be a suitable option for our study. Instead, variation of RAs within sociology will be explored by comparing RAs across sociology journals. Since RAs not only reflect the values and norms of an academic community but are also shaped and negotiated to meet quality requirements (i.e. theoretical, paradigmatic, methodological, textual and language-related) set by journal editors, reviewers and publishers (Swales 1990), there is reasonable ground to expect that this approach could reveal aspects of intra-disciplinary variation.

Our study draws on data from six small corpora of RAs, each compiled from a different high impact journal. All the journals are indexed in the SSCI under the category of sociology. To balance the prevalence of journals of Anglo-American origin in the SSCI, half of the selected journals were randomly selected from non-English speaking countries. All the articles included in the corpora are written in English. To provide further categorization of the journals, both the journal descriptions and expert informants were consulted. The journals' positioning statements did provide information about the niche the journal is intended to occupy; however, the niche was either too vague or too complex relative to the overlapping fields internal and external to sociology, theoretical approaches and methodologies. As the positioning of journals is not static but changes over time, further categorizations of the selected journals were felt to be problematic. We could only establish that three of the journals have a narrower topical focus and three wider or more general.

The most recent RAs accessible in digital format were those for the year 2006. All RAs published in that year were included. Two of the journal corpora had to be expanded with further volumes to ensure greater representativeness of the individual corpora. Whole texts were used for the analysis, but they exclude data about authors' affiliation, footnote descriptions of research projects, as well as acknowledgements, references

and various diagrams. All in all, the six corpora (Table 1) comprise 245 RAs totalling 2,099,276 words.

This study, however, is not only corpus-based, but also corpus-driven. Gledhill (1995, 2000) has shown that the analysis of salient grammatical words and their collocations reveals the generic and stylistic characteristics of RAs, a finding which was confirmed by Groom (2010). Scott (2001) further found that key grammatical words can also reveal ideational and interpersonal characteristics of texts. It is thus hoped that they can also provide a reliable means to uncover intra-disciplinary variation within sociology RAs.

| Journal | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| *American Journal of Sociology* (AJS) | USA | 1895 | w. | 2006 | 533,032 | 35 | 13,326 | 5,486 |
| *Journal of Marriage and Family* (JMF) | GB | 1939 | n. | 2006 | 633,881 | 84 | 7,546 | 1,918 |
| *Social Forces* (SOF) | USA | 1922 | w. | 2006 | 307,900 | 41 | 7,510 | 1,560 |
| *Development and Society* (DAS) | KR | 1971 | n. | 2006/ 2007 | 164,760 | 26 | 6,337 | 2,033 |
| *Demographic Research* (DER) | D | 1999 | n. | 2006 | 240,692 | 32 | 7,522 | 3,132 |
| *Sociologický časopis* (SCA) | CZ | 1965/ 1993* | w. | 2006/ 2007/ 2008 | 219,011 | 27 | 8,112 | 1,510 |

**Table 1. Overall characteristics of the sociology journal corpora.**

Note: 1: Origin, 2: Established, 3: Focus, w.: wider, n.: narrower, 4: Volume included in the corpus, 5: Number of words, 6: Number of RAs, 7: Length of RAs (mean), 8: Standard deviation
* The first English edition was published in 1993

Following Gledhill's (1995, 2000) and Scott's (2001) methodology, grammatical keywords were first elicited from all the corpora together by comparing them with the written component of the British National Corpus (BNC) using log likelihood (LL) statistics. Among the key grammatical words, the two most salient were selected for further analysis:

*among* and *between*. These words were also identified as keywords when the individual corpora were compared to the written component of the BNC. Next, the frequencies of the selected keywords were compared across the journal corpora in order to establish whether statistically relevant differences exist among them. The words were then studied in the contexts of the individual journals. WordSmith Tools 3 (Scott 1998) and WordSmith Tools 5 (Scott 2008) were used for the major part of the analysis. The online log likelihood calculator of the University of Lancaster was used to compare figures related to the principal uses of individual keywords. There are some differences in the way log likelihood is calculated by WordSmith Tools and the online log likelihood calculator; nevertheless, at a p-value threshold of 0.001 these seem of minor importance.

## 4. Findings

First, the length of RAs was compared. The mean values of article length are similar in three of the journals, JMF, SOF and DER, albeit with varying standard deviations (Table 1). The mean length of RAs in SCA is somewhat higher and that in DAS is somewhat lower. Interestingly, RAs are not only longest in AJS, but also over 75% of the articles published in this journal are longer than the great majority of those published in the other journals (Figure 1).

*Among* and *between* are among the least frequent prepositions in English, yet they are significantly more frequent in academic language (Biber et al. 1999). Table 2 shows the normalized frequencies of the selected keywords per 100,000 words. By using log-likelihood statistics (the p-value threshold was set to 0.001), significant differences were found among some of the corpora (Table 3). These are discussed in detail further below.

As opposed to bound prepositions, *among* and *between* are classified within the group of free prepositions (Biber et al. 1999), which means that they have an independent meaning. Quirk et al. (1985) distinguish between spatial, temporal and metaphorical or abstract use of the prepositions *among* and *between*. The abstract use of *between* refers to relationships, contrast and affinity between discrete objects, while *among* refers to non-discrete objects. Groom (2007) identifies different groups of semantic sequences in which these two words participate and observes that *among* serves to describe a phenomenon within a social group or to evaluate a member of a group. *Between* has a similar function, but as

opposed to *among*, it sets the limits to a phenomenon, conveys semantic associations of relationships, divisions and oscillations.

In sociology RAs, *among* and *between* typically introduce metaphorical or abstract relations. Rather than pointing to spatial information, they predominantly participate in the descriptions of social phenomena and methodology.

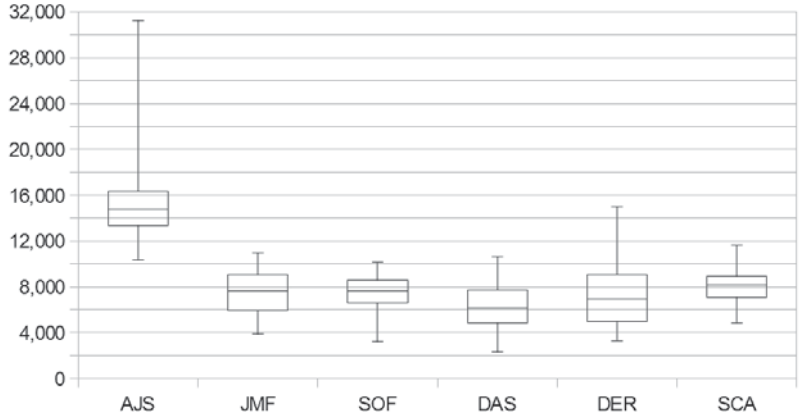



**Figure 1. Variability of RA length.**

| | AJS | JMF | SOF | DAS | DER | SCA |
|---|---|---|---|---|---|---|
| *among* | 96 | 6 | 143 | 144 | 163 | 94 |
| *between* | 188 | 346 | 269 | 295 | 329 | 199 |

**Table 2. Normalized frequencies per 100,000 words.**

| | ***among*** | ***between*** |
|---|---|---|
| JMF:AJS | +81.2 | +275.2 |
| SOF:AJS | +36.1 | +58.3 |
| DAS:AJS | +24.6 | +63.3 |
| DER:AJS | +59.1 | +135.2 |
| SCA:AJS | | +21.2 |
| SOF:JMF | | -39.5 |
| DAS:JMF | | |
| DER:JMF | | |
| SCA:JMF | -26.1 | -60.0 |
| DAS:SOF | | +16.3 |
| DER:SOF | | |
| SCA:SOF | -11.6 | |

| | | |
|---|---|---|
| DER:DAS | | |
| SCA:DAS | | |
| DER:SCA | +24.5 | +31.3 |

**Table 3. Comparisons of *among* and *between* across the corpora.**

## 4.1. *Among*

*Among* is used in 237 RAs (97%): it is not used in 20% of articles in DER, 6% of articles in JMF and 5% of articles in SCA. The content analysis of the RAs in which *among* was not used shows that this could be related to a number of elements: the author's choice of individuals as the unit of analysis rather than groups of indiscrete entities, the research methods chosen (e.g. an interview), the topic (e.g. analysis of reliability of measurement scales), or a combination of these factors. In all corpora, however, there are more exceptionally small frequencies of *among* than exceptionally large.

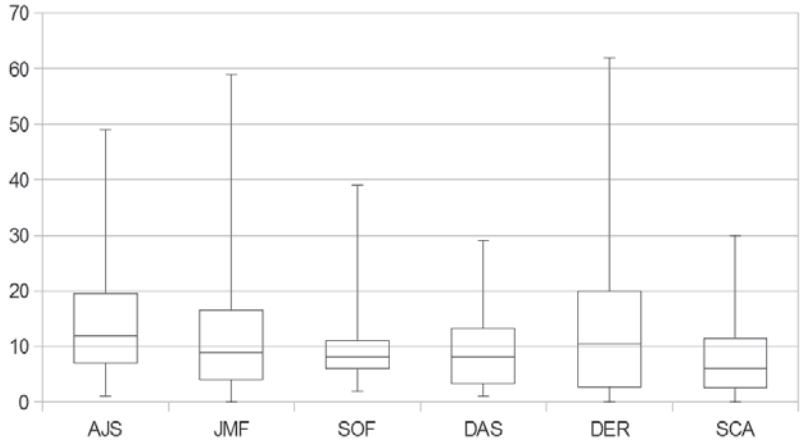



**Figure 2. Distribution of *among* in the six journal corpora.**

The comparisons across the corpora (Table 3) reveal that *among* is significantly less frequent in AJS than in the other corpora. The reason for this may lie in a more general and theoretical orientation of the journal. The journal's scope suggests that AJS publishes articles in which authors aim at developing theories, setting conceptual frameworks and discussing

innovative methods rather than reporting on results of empirical research. Curiously, *among* is also significantly less frequent in SCA, which is not theoretically oriented. Its scope is rather fuzzy and it mirrors the journal's endeavours to survive in a highly competitive publishing environment by broadening its initially more regional perspective to "every area of sociology". The content analysis of the texts shows that relatively few articles in this journal actually investigate relations between social groups. In line with the journal's orientation, most of the articles deal with issues related to social and political development in post-communist societies, but there are also some articles that are theoretical in nature.

In sociology RAs, *among* tends to be used when phenomena or relations in social groups or in groups of phenomena taken as entities are described. In all the corpora, it is mainly used in the pattern Noun + *among* + Plural Noun with the semantic association

SOCIAL PHENOMENON + *among* + SOCIAL GROUP

– the latter representing the sample population. *Among* thus allows the author to relate the observed social phenomena to a selected sample in the study, to analyse the variables, and to report on the results of research.

**(1)**
The differences in leisure patterns *among* men and women are more contextual than biological. (DAS06_10)

In the example above, the author actually observes a phenomenon in two distinct social groups (i.e. the group of men and women) and compares these two groups by explaining the variables. *Among* occasionally relates social phenomena not only to the social groups investigated, but also to other types of groups; for example, to a research community.

**(2)**
It has long been axiomatic among sociologists of work that even seemingly powerless groups can find ways of acting back on their presumptive superiors, thereby shaping their work situations in accordance with their own needs (Mechanic 1962; Halle 1984; Simpson 1989). Curiously, however… (AJS06_14)

In the latter case, *among* introduces a broader context in which the attitude of the research community towards a phenomenon is introduced, or the author's agreement or disagreement with a certain position or theory

is expressed, previous research is related to, or a theoretical context for research is established. In articles where this function was identified greater emphasis is placed on conceptualisation. The function seems to be more marked in journals which are more theoretically and narratively oriented (i.e. AJS, SOF).

The author's choice of the group under investigation closely corresponds to the niche of the journal. This correspondence is seen especially in the journals with a narrower focus: JMF, DER and DAS. The majority of the samples in the JMF thus relate to groups of family members and to family types while the samples in DAS refer to economic actors. But as even the more narrowly focused journals allow for some degree of interdisciplinarity, the observed social group could also be one that is less typical of a journal. Furthermore, the author's choice of the group under investigation also seems to be related to the geographic coverage of the journal: regional or global. The adjectives of nationality which sometimes modify the noun head denoting the social group are informative about the geographic coverage of research published in the journals. Thus, the regional (Non-Anglophone) journals tend to publish nationally or regionally relevant research while in the global (Anglophone) journals UK/USA based research or international comparative studies predominate.

Higher frequencies of *among* are found in RAs where the sample population is a unit of indiscrete individuals or where subgroups of the sample population are described, especially when these exhibit particular characteristics (example 3 below). In the latter case, the author compares phenomena and trends across a number of social groups.

**(3)**
A rich literature exists concerning tool use among birds, most often among corvids (crows, jay, ravens and jackdaws). (SOF06111)

As regards the phenomena described in the pattern Noun + *among* + Plural Noun, Francis, Manning and Hunston (1998) distinguish 10 meaning groups of nouns. In our corpora, only three meaning groups were found that are shared by at least three of the corpora: "the relationship group", "the differences group" and "the conflict group". We will focus on these nominal groups only. Their frequencies were compared across the corpora (see Table 4) using log likelihood. Differences at the cut-off point for statistical significance of $p<0.01$ (LL=6.63) are also displayed to illustrate stronger trends.

The "differences group" refers to differences or similarities within the observed entities. In all six corpora, differences rather than similarities are highlighted. Differences in sociology are often perceived as negative,

directly leading to social inequality. This group is significantly less frequent in SCA than in the other corpora.

The "conflict group" refers to fight, argument or contest (see example 4). This group appears to be significantly more frequent in AJS and to a degree in DAS. Social conflicts are linked with social inequalities, competition, power relations and dominance. They are considered to be a threat to public safety and to the stability of a society.

| | Nominal patterns | | | Verbal patterns |
|---|---|---|---|---|
| | Relationship | Difference | Conflict | |
| JMF:AJS | | - 6.78 | - 9.92 | + 12.05 |
| SOF:AJS | | | - 11.80 | + 20.49 |
| DAS:AJS | + 28.82 | | | |
| DER:AJS | | | - 17.89 | + 21.78 |
| SCA:AJS | | - 26.16 | - 16.52 | + 13.22 |
| SOF:JMF | | | | |
| DAS:JMF | + 22.83 | + 14.36 | | - 15.64 |
| DER:JMF | | | | |
| SCA:JMF | | - 21.96 | | |
| DAS:SOF | + 7.89 | | | - 22.54 |
| DER:SOF | - 11.32 | | | |
| SCA:SOF | - 7.98 | - 18.27 | | |
| DER:DAS | - 30.74 | | - 10.81 | + 24.00 |
| SCA:DAS | - 25.26 | - 33.82 | - 10.15 | + 17.73 |
| DER:SCA | | + 18.12 | | |

**Table 4. Comparisons of nominal and verbal patterns of *among* across the corpora.**

**(4)**
Legal institutions, including administrative agencies and courts, are important sites of political conflict among challengers, dominant social groups, and the state (Bernstein 2001; Handler 1978; Pedriana 2004; Pedriana and Stryker 2004). (AJS06_16)

Francis, Manning and Hunston's (1998) "communication group", originally referring to communication and transactions between people and groups, was here renamed "relationships group" to highlight the prevailing meaning associations found in our corpora. Nouns that describe the nature of links between different variables (see example 5) are also placed in this group. As opposed to the "conflict group" and the "differences group", this group of nouns points to the homogeneity of social groups and to cohesiveness in society. Nouns from the "relationship group" are significantly more frequent in DAS than in the other corpora.

**(5)**
The cointegration implies a long run relationship among variables. (DAS06_11)

Other relevant patterns with *among* found in the sociology journal corpora are patterns with verbs: Verb + *among* + Plural Noun and Verb + Adjective + *among* + Plural Noun. The log likelihood data across the corpora show that patterns with verbs are significantly less frequent in AJS and DAS than in the other corpora. Interestingly, these two corpora display significantly higher frequencies of nominal patterns co-occurring with *among*: the former with the semantic association of conflict and the latter with the semantic association of relationship. This is not surprising as the verbal patterns seem to be used with a different communicative purpose. They tend to set a phenomenon in context (see example 6), identify a member (or a subgroup) pertaining to a larger group (see example 7), and to evaluate this member or a subgroup (see example 8). The verbs related to *among* predominantly fall within the "occur" group of verbs. The adjectives following the verb and preceding *among* mostly evaluate the frequency of a phenomenon in the context of research.

**(6)**
Stylistic differences in whale song occur among different orca populations (Randell and Whitehead 2001). (SOF06111)

**(7)**
These analyses are among the first to measure the distribution of various forms of child-care assistance received by families with young children. (JMF06_01)

**(8)**
Single-year and consistent earnings advantages are more common among Black wives than White wives, by both measures. (JMF06_73)

Structures with *among* that are in clause initial position seem to provide topical orientation to the clause. They participate in the description of the selected group and at the same time set the grounds for a comparison with other groups, variables or results (see example 9).

**(9)**
There is greater material risk among Black children living with cohabiting rather than married biological parents, which is explained by parent's education. Among Hispanic children, we do not observe marital status differences in high material risk. (JMF06_03)

The comparison across the corpora points to a relatively higher frequency of clause initial structures in JMF when compared to AJS, DER and SCA (log likelihood values were 9.74, 12.63 and 10.49, respectively). It seems that authors use clause initial structures with *among* in JMF to thematise different ethnic, religious or other types of social groups observed in order to present the differences among them.

All in all, the results of the comparisons show that the overall frequency of *among* is significantly lower in AJS compared to all the other corpora; a more detailed analysis of the patterns *among* is used in reveals differences across the corpora that seem to be related to the topical and methodological orientations of journals.

## 4.2. *Between*

In contrast to *among*, *between* was present in all articles included in our corpora. Its frequencies were consistently higher than those for *among*, which is in line with Biber et al. (1999). Slightly more authors used *between* more frequently in AJS. Although the spread of *between* and the range of its use is quite similar in JMF, too, there are more of those who use *between* less frequently in this journal. The use of *between* is most consistent across the RAs in DAS.

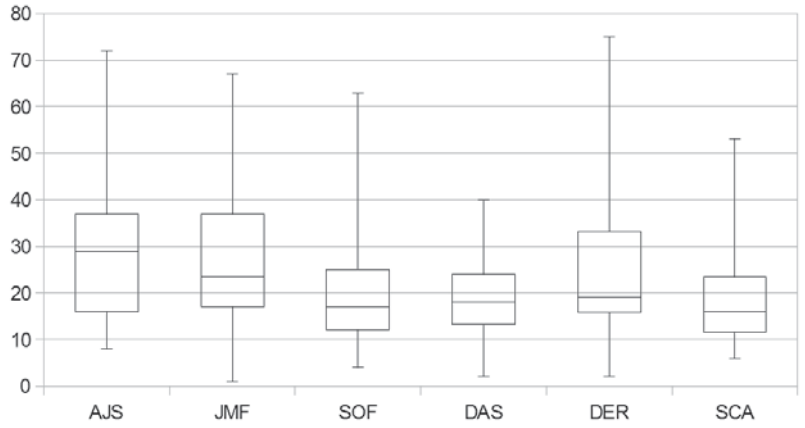



**Figure 3. Distribution of *between* in the six journal corpora.**

*Between* has two main functions in sociology RAs: firstly, it relates a phenomenon to two or several observed groups taken as discrete entities, thus forming the following pattern:

PHENOMENON + *between* + PHENOMENA

There are two or more phenomena or social groups included in the phenomena described (see example 10). There is another pattern with *between* that introduces two extreme values which impose the limits on the conditions of the observed phenomenon or set a time frame (see example 11):

*between* + LIMIT + *and* + LIMIT

**(10)**
Ties between foreign-owned and domestically owned firms are just as likely as ties within these segments. (AJS06_08)

**(11)**
The trends in academic research are similar to those in higher education; for instance, the number of specialized economic reviews worldwide increased five times between 1959 and 1993, from about 500 to over 2,500 […]. (AJS06_22)

The comparison of frequencies of the preposition *between* shows trends that are similar to those observed with *among*: *between* is used significantly less often in AJS than in the other corpora (cf. Table 3). *Between* seems to be less frequent in RAs in which authors discuss formal theories, theoretical models and the development of research tools and more frequent in empirical RAs researching relations between social groups (e.g. JMF).

Francis et al. (1998) describe two types of patterns with *between*: Adjective *between* Plural Noun and Noun *between* Plural Noun. While the former pattern was quite rare in our corpora, the latter proved to be predominant. Of the 15 different meaning groups associated with the second pattern by them, four were not found in our specialised corpora. We also added a new group that was not included in Francis et al., but was present in our corpora; i.e. the "periods and other types of ranges" group, also found by Lindstromberg (2010).

Some of the meaning groups show very low frequencies, others are not used in all corpora. The most frequent meaning groups were: the "periods and other types of ranges group", the "relationship group", the "difference

group" the "interaction group", the "similarity group", and the "fight group". It seems that *between* typically introduces ideas of time-setting, value description, relations, and differences between social agents and phenomena. These meaning groups were compared across the corpora using log likelihood (Table 5). To highlight stronger trends in the ways *between* is used, the table also shows differences at the statistical significance level of $p<0.01$.

| | Periods and other types of ranges | Relationships | Interaction | Differences |
|---|---|---|---|---|
| JMF:AJS | + 18.15 | + 257.92 | | + 93.60 |
| SOF:AJS | | + 30.74 | + 42.67 | +23.04 |
| DAS:AJS | | + 24.73 | | + 15.96 |
| DER:AJS | + 13.64 | + 22.71 | | + 111.88 |
| SCA:AJS | + 34.31 | | - 8.87 | + 30.43 |
| SOF:JMF | | - 71.91 | + 21.24 | |
| DAS:JMF | - 13.74 | - 34.51 | | - 6.85 |
| DER:JMF | | - 58.98 | | + 8.75 |
| SCA:JMF | + 7.14 | - 124.23 | - 11.06 | |
| DAS:SOF | - 7.38 | | - 22.29 | |
| DER:SOF | | | - 11.82 | + 7.30 |
| SCA:SOF | + 9.91 | - 12.39 | - 40.04 | |
| DER:DAS | + 12.73 | | | + 19.65 |
| SCA:DAS | + 26.46 | - 14.16 | | |
| DER:SCA | | + 11.55 | + 9.27 | |

**Table 5. Comparison of meaning groups in patterns with *between*.**

The nouns associated with "periods and other types of ranges" were significantly more frequent in SCA and significantly less so in DAS compared to the other journals. This pattern introduces two extreme points which delimit periods or other types of intervals related to the description of results. In all journals, the overwhelming majority of nouns in this group refer to time periods. The fact that this meaning group is significantly more frequent and that the other meaning groups in this pattern are significantly less frequent in SCA than in the other corpora seem to point to the journal's preoccupation with the period of transition in Central European post-communist states and with the tendency to identify the observed phenomena from a chronological perspective.

The "relationships group" was the largest meaning group participating in the pattern Noun *between* Plural Noun. The data point to clear trends in three of the journals: AJS and SCA displayed significantly lower

frequencies of nouns used in this group, while JMF showed significantly more frequent use of this group compared to the other journals. In JMF, *between* thus participates in the description of close relations, relations between family members, and relations between family members and social groups. It also reveals a pronounced concern to determine the complex nature of the links between dependent and independent variables or sets of variables which exert an influence on the observed phenomena. In some cases the hypothesis or the research focus itself revolves around the question of the association between distinct social phenomena and the question of the possibility to further examine, test, evaluate or explain the nature of this association or its absence.

The "differences group" is the second most frequent meaning group in our corpora. If the "differences group" with *among* was often accompanied with the idea of conflict, *between* establishes the perspective of differentiation, search for clear identification of a group and the distinction between groups and concepts. The comparisons of the frequencies of this group show that it is the least frequent in AJS. The low frequency of this meaning group could be explained by the theoretical orientation of the journal. On the other hand, DER shows tendencies towards higher frequencies of the use of this meaning group compared to JMF, SOF and DAS. In DER, the "differences group" is not used only to describe the differences between social groups and variance across variables, but also to introduce relations of comparative subtraction in mathematical functions as part of statistical and mathematical modelling in demographic research, which is the distinguishing characteristic of this journal (see example 12).

**(12)**
In Figure 5 the ordinate (set off along the y-axis) is the difference between (i) the CFR for the ever-married in an educational group and (ii) the corresponding CFR for the never-married in the same group. (DER06_16)

The "interaction group" is significantly more frequent in SOF. This group is predominantly linked with questions of methodology, statistical modelling − e.g. regression and analytic techniques (see example 13). Only a small number of examples in this group account for social interaction. However, it is interesting to note that the share of examples pointing to social interaction is higher in those journals which exhibit considerably lower frequencies of nouns from the "interaction group" (i.e. SCA and DAS). The tendency to use this meaning group more frequently therefore points to stronger methodological considerations of a journal.

**(13)**
The interaction between the intercept for maternal depression and family instability, for example, tested whether the mean level of maternal depression moderated the link between family instability and child behavior. (SOF06309)

In comparing Tables 3 and 5, we can conclude that the statistically significant differences in the frequencies of *between* across the corpora do not always correlate with the statistically significant differences in the frequencies of the meaning groups associated with *between*. On the one hand, we can see that both the overall frequency of *between* and most of the meaning groups associated with *between*, i.e. the "relationships", "differences" and also partly the "periods and other types of ranges" groups, are significantly lower in AJS than in the other corpora. On the other hand, while the discrepancy in the frequency of *between* in SCA and SOF is insignificant (log likelihood was -3.98 at $p< 0.05$), the comparison of the meaning groups associated with *between* reveals that, when compared to SCA, *between* is used significantly more frequently in SOF with nouns belonging to the meaning groups of "interactions" and "relationships" and relatively less frequently with those belonging to the "periods and other types of ranges" group. Therefore, significant difference in semantic associations of a word may exist even where the difference in the overall word frequency is not significant.

## 5. Discussion and conclusion

In this chapter we set out to establish whether linguistic variation exists within the field of sociology. According to sociologists, their discipline lacks a paradigmatic, theoretical and methodological core and it is affected by increasing fragmentation and segmentation as a result. As they tend to gather around research topics and publish their research in specific journals where RAs share certain theoretical and methodological approaches, it seemed that comparing corpora, each representing a particular journal, might reveal intradisciplinary variation if it existed.

From the analysis of the six sociology journal corpora it emerges that *among* and *between* are not only the most salient but are also of significance in each corpus. What is more, the corpora displayed certain differences in the frequencies of these two prepositions: some of them were significant.

Both *among* and *between* were used in a smaller number of different patterns. The nouns they combined with in these patterns were extremely diverse, though they shared some semantic associations. Based on these,

meaning groups (Francis, Hunston and Manning 1998) were formed and their use in the corpora analysed. The results revealed some of the nuances of intradisciplinary variation.

Both *among* and *between* seem to be significantly less frequently used in journals with a more theoretical orientation (i.e. focusing more on social phenomena and their conceptualisation than on relations among social groups). *Among* is nonetheless used more often in theoretical journals to relate authors' views and findings to those of their academic community.

Authors thus tend to use *among* when social phenomena are described in the context of social groups, or when relations among groups or variables are compared and discussed. Verbal patterns with *among* suggest different communicative purposes: identification and evaluation of groups observed, placing phenomena in context and topical orientation of the reader.

Some of the semantic associations of the nominal and verbal patterns with *among* and *between* were found to be more frequent in some of the journals. The various combinations of distinct semantic associations seem to form a blend that is specific to individual journals.

*Among* and *between* allow the author to identify, select and highlight the objects of research, to relate them to prominent social actors and present them either as phenomena or as objects. They also participate in patterns that are related to methodological issues. Occasionally, they participate in the topical orientation of the content. The relations that these two prepositions help to express are deeply embedded in the sociological perspective: sociology studies human society and it is therefore interested in the functioning of social groups, the interactions between them, and in the underlying patters formed by them. Due to its multi-paradigmatic nature, sociology is also very much dependent on conceptualisations of social phenomena and on the justification of its methodology. These two prepositions are therefore related to the core of sociological concern.

This corpus-based study confirms that there is intradisciplinary variation in sociological RAs that is not due to the stylistic flexibility with which authors use language. The differences between the preferred meanings and values found in our journal corpora could be attributed to differences in the research focus and theoretical positionings of the authors, the methodologies they used, and also the niches occupied by the journals. Gledhill's (1995, 2000) and Scott's (2001) corpus-driven methodology successfully elicited two prepositions that not only reveal what sociology RAs are about but also uncover some aspects of the intradisciplinary variation within the field. The present study also shows that significant differences in frequencies of words across corpora may or

may not be maintained when the semantic associations of the word are compared and that there may be significant discrepancies between semantic associations of words even where no significant difference is identified on the word frequency level. Corpus-based and corpus-driven analysis therefore proved to be a fruitful approach.

However, our study suffers from a number of limitations. First of all, due to the time-consuming nature of this approach, only two keywords were explored in the context of sociological RAs. To do intradisciplinary variation justice, more keywords would have to be investigated. Next, the analysis is mainly limited to the predominant collocates of *among* and *between* and the patterns of semantic association that authors are primed to expect and use. Less frequent patterns and other lexical primings remain unaccounted for. Finally, it was felt that the DAS corpus was perhaps too small for the investigation of the prepositions that are not among the most common. Since most of the journal corpora included only one volume of RAs, it is unclear whether a larger volume of RAs would evoke different aspects of sociological content. Although journals change over time due to changing research trends, editorial policies, publication practices, and so on, some of their characteristics remain stable.

It was felt that future corpus studies of intradiciplinary and cross-disciplinary variation in academic discourse would benefit greatly from improved corpus selection criteria. Randomly chosen RAs from a handful of different journals may provide a limited view of a discipline and overlook intradisciplinary variation. Even though our study has clearly confirmed intradisciplinary variation in the field of sociology, further studies would be needed to uncover the complex interplay of factors behind intradisciplinary variation.

## References

Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad and Edward Finegan. 1999. *Longman grammar of spoken and written English*. Harlow: Longman.

Bondi, Marina. 2009a. Polyphony in academic discourse: A cross-cultural perspective? In *Cross-cultural and cross-linguistic perspectives on academic discourse*, ed. Eija Suomela-Salmi and Fred Dervin, 83-108. Amsterdam: John Benjamins.

—. 2009b. In the wake of the terror: Phraseological tools of time setting in the narrative of history.In *Academic writing: At the interface of corpus and discourse*, ed. Maggie Charles, Diane Pecorari and Susan Hunston, 73-90. London: Continuum.

Brett, Paul. 1994. A genre analysis of the result sections of sociology articles. *English for Specific Purposes* 13(1): 47-56.

Bruce, Ian. 2009. Results sections in sociology and organic chemistry articles: A genre analysis. *English for Specific Purposes* 28(2): 105-124.

Charles, Maggie. 2007. Argument or evidence? Disciplinary variation in the use of the Noun that pattern in stance construction. *English for Specific Purposes* 26(2): 203-218.

Dahl, Trine. 2009. The linguistic representation of rhetorical function: A study of how economists present their knowledge claims. *Written Communication* 26(4): 370-391.

Dogan, Mattei. 2000. Sociology among the social sciences. In *Encyclopedia of sociology*, Volume 5, ed. Edgar F. Borgatta and Rhonda J.V. Montgomery, 2913-2926. New York: Macmillan.

Francis, Gill, Susan Hunston and Elisabeth Manning. 1996. *Collins COBUILD grammar patterns: Verbs.* London: HarperCollins.

—. 1998. *Collins COBUILD grammar patterns: Nouns and adjectives*. London: HarperCollins.

Gledhill, Christopher 1995. Collocation and genre analysis: The phraseology of grammatical items in cancer research abstracts and articles. *Zeitschrift für Anglistik und Amerikanistik* 43(1): 11-36.

—. 2000. *Collocations in science writing*. Tübingen: Gunter Narr.

Groom, Nicholas. 2007. *Phraseology and epistemology in humanities writing: A corpus-driven study*. Ph.D. Thesis. University of Birmingham.

Harwood, Nigel. 2009. An interview-based study of the functions of citations in academic writing across two disciplines. *Journal of Pragmatics* 41(3): 497-518.

Hoey, Michael. 2005. *Lexical priming: A new theory of words and language*. London: Routledge.

Holmes, Richard. 1997. Genre analysis, and the social sciences: An investigation of the structure of research article discussion sections in three disciplines. *English for Specific Purpose*s 16(4): 321-337.

Hyland, Ken. 1999. Academic attribution: Citation and the construction of disciplinary knowledge. *Applied Linguistics* 20(3): 341-367.

—. 2001. Humble servants of the discipline? Self-mention in research articles. *English for Specific Purposes* 20(3): 207-226.

—. 2006. Disciplinary differences: Language variation in academic discourses. In *Academic discourse across disciplines*, ed. Ken Hyland and Marina Bondi, 17-45. Bern: Peter Lang.

—. 2007. Applying a gloss: Exemplifying and reformulating in academic discourse. *Applied Linguistics* 28(2): 266-285.
—. 2008. Disciplinary voices: Interactions in research writing. *Journal of English Text Construction* 1(1): 5-22.
Hyland, Ken, and Françoise Salager-Meyer. 2008. Scientific writing. *Annual Review of Information Science and Technology* 42(1): 297-338.
Jarc, Mojca and Šarolta Godnič Vičič. 2012. The long and winding road to international academic recognition: The case of Slovene social sciences authors. In *Akademski jeziki v času globalizacije / Academic languages in the era of globalisation*, ed. Sonja Starc, 229-241. Koper: Annales.
Kanoksilapathan, Budsaba. 2007. Rhetorical moves in biochemistry research articles. In *Discourse on the move: Using corpus analysis to describe discourse structure*, ed. Douglas Biber, Ulla Connor and Thomas A. Upton, 73-119. Amsterdam: John Benjamins.
Lorés Sanz, Rosa. 2004. On RA abstracts: from rhetorical structure to thematic organisation. *English for Specific Purposes* 23(3): 280-302.
—. 2011. The construction of the author's voice in academic writing: The interplay of cultural and disciplinary factors. *Text & Talk* 31-2: 173-193.
Luzón Marco, Maria José. 2000. Collocational frameworks in medical research papers: A genre-based study. *English for Specific Purposes* 19(1): 63-86.
Mur-Dueñas, Pilar. 2011. An intercultural analysis of metadiscourse features in research articles written in English and in Spanish. *Journal of Pragmatics* 43(12): 3068-3079.
Ozturk, Ismet. 2007. The textual organisation of research article introductions in applied linguistics: Variability within a single discipline. *English for Specific Purposes* 26(1): 25-38.
Pontille, David. 2003. Authorship practices and institutional contexts in sociology: Elements for a comparison of the United States and France. *Science, Technology, & Human Values* 28(2): 217-243.
Posteguillo, Santiago. 1999. The schematic structure of computer science research srticles. *English for Specific Purposes* 18(2): 139-160.
Quah, Stella R., and Arnaud Sales. 2000. On consensus, tensions and sociology at the dawn of the 21st century. In *The international handbook of sociology*, ed. Stella R. Quah and Arnaud Sales, 1-32. London: Sage.
Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik. 1985. *A comprehensive grammar of the English language*. London: Longman.

Ruiying, Yang and Desmond Allison. 2003. Research articles in applied linguistics: Moving from results to conclusions. *English for Specific Purposes* 22(4): 365-385.

Salager-Meyer, Françoise, María Ángeles Alcaraz Ariza, and Nahirana Zambrano. 2003. The scimitar, the dagger and the glove: Intercultural differences in the rhetoric of criticism in Spanish, French and English medical discourse (1930-1995). *English for Specific Purposes* 22(3): 223-247.

Samraj, Betty. 2005. An exploration of a genre set: Research article abstracts and introductions in two disciplines. *English for Specific Purposes* 24(2): 141-156.

Scott, Mike. 1998. *WordSmith Tools 3*. Oxford: Oxford University Press.

—. 2001. Comparing corpora and identifying key words, collocations, frequency distributions through the WordSmith Tools suite of computer programs. In *Small corpus studies and ELT*, ed. Mohsen Ghadessy, Alex Henry and Robert L. Roseberry, 47-67. Amsterdam: John Benjamins.

—. 2008. *WordSmith Tools 5*. Liverpool: Lexical Analysis Software.

Swales, John. 1990. *Genre analysis: English in academic and research settings*. Cambridge: Cambridge University Press.

Testa, James. 2011. *The globalisation of web of science: 2005-2010*. Philadelphia/London: Thomson Reuters.

Vanderstraeten, Raf. 2010. Scientific communication: Sociology journals and publication practices. *Sociology* 44(3): 559-576.

# CHAPTER FOUR

# KNOWLEDGE CONSTRUCTION AND KNOWLEDGE PROMOTION IN ACADEMIC COMMUNICATION: THE CASE OF RESEARCH ARTICLE ABSTRACTS—A CORPUS-BASED STUDY

MICHELE SALA
UNIVERSITÀ DI BERGAMO, ITALY

## 1. Introduction

The research article abstract (RAAB) is a highly standardized genre (Salager-Meyer 1990; Cross and Oppenheim 2006; Swales and Feak 2009) pointing to another self-standing text unit representative of a recognizable genre, that is the research article (RA), which the RAAB is meant to reflect both at the representational level − by introducing the main theme and problematic aspects which are going to be assessed in the abstracted text − and at the rhetorical and cognitive level − by mirroring the argumentative or expository organisation of the ensuing RA (i.e. in gap-filler, problem-solution, or topic-method sequences) and introducing the relevant research parameters (i.e. study design, research questions and aim, data analysis, etc.) through conventionalized moves and rhetorical strategies distinctive of the relevant disciplinary domain. In this sense RAABs are a miniature version of the corresponding RA − although with a marked descriptive character (due to space constraints) and promotional function (being intended to elicit the reader's interest, cf. Hyland 2000) − which help establish and corroborate the epistemology of a given discipline by variably codifying objectivity and meaning negotiability, assertiveness and defensiveness, authoritativeness and cooperation.

From a generic point of view, there is general agreement (Hyland 2000; Swales and Feak 2009; Bondi 2010; Bondi and Cavalieri 2012) as to the RAAB textual structure reflecting a standard template consisting of four moves meant to cognitively organise the knowledge content of the associated RA − namely by introducing the purpose, the methodology, the main results and, possibly, their discussion (Bhatia 1993) − and which are usually prefaced by an introductory move (traditionally labelled "background") with the function of situating the research activity within a recognizable epistemological framework (Dos Santos 1996), i.e. the disciplinary and sub-disciplinary scenario, a given research tradition, etc. Whereas the moves corresponding to the presentation and the discussion of the results can be missing − whereby it is possible to distinguish between informative and indicative RAABs, the latter type only presenting the scope and the purpose of the research but not the results (Lorés Sanz 2004) − background elements tend to be consistently lexicalized, and have increasingly done so in the course of the last decade (Bondi and Cavalieri 2012), either in a self-standing initial section or scattered throughout the body of the RAAB. This may be due to the promotional character which is typical of the genre, in that the RAAB needs to be (or be perceived) as transparent, meaningful and disciplinary relevant by the targeted audience for it to be easily interpreted and perlocutionarily lead to reading the associated RA. For this reason RAABs in all disciplines tend to include at least a step, if not a whole move, reflecting the CARS model found in RA introduction (Swales 1990), which is meant to establish a territory and a niche within it which will then be occupied by the informative material provided in the main text. In other words, this step is intended to "set the scene" by highlighting a knowledge gap that is disciplinary relevant and which motivates and justifies the research activity reported in the RAAB. The background has the cognitive and rhetorical function of providing the basis on which to substantiate knowledge claims, to present the scenario with respect to which new information acquires relevance, and to establish the scientific tradition within which the study at hand will have to be set and interpreted. By exploiting this function, writers manifest their competence in the field they are dealing with, portraying themselves as competent community members and, as a consequence, marking their activity as the work of an expert.

For these reasons RAABs have been claimed to be a "representation" of the associated RA (Bazerman 1988; Berkenkotter and Huckin 1995; Bondi 1997), in that authors anticipate not only the content but also, and most distinctively, the way content is going to be assessed in the RA, both at an organisational level and at a meta-textual level, by "metadiscursive

references to the original article and its procedures" (Bondi and Cavalieri 2012, 44). This way of organizing the material, as well, may have a promotional character. As a matter of fact, the adherence to codified and recognizable macro-textual standards reveals the author as a competent "writer", someone aware of the appropriate way to conceptualize, discuss and report discipline-related objects, activities of knowledge production, processes and roles.

The competence and expertise of the writer can also be displayed at a micro-linguistic level by the consistent resorting to discipline-specific lexis and phraseology, through the use of which the targeted audience is given keywords whose denotation and, especially, discipline-specific conventional use is immediately transparent and meaningful. This is particularly relevant when such key terms are used to thematize the activity of knowledge-making within the discipline, that is to say, to introduce and explain contents, methods, procedures and results. Indeed, due to the specificity of the knowledge object and the processes of its dissemination, the various disciplines tend to resort to specific and standardized linguistic resources to represent such processes as acts of either research, cognition or argumentation (Biber et al. 1999; Hyland 2002). Since scientific activity is at the same time an act of observation, interpretation of the observed phenomena and discussion of their meaning, the aim of this analysis is to see to which knowledge-making paradigm scholars preferably associate their work when writing about it, and especially when abstracting it in RAABs. A second line of investigation will then compare such knowledge frameworks as they are used to provide background information, on the one hand, and introduce new informative material, on the other (the latter being the core of the abstracted RA). This is particularly relevant since, as we will see, consistency in the use of such resources would be a sign of continuity between existing and new knowledge-making practices, whereas variation would indicate a paradigm shift in conceptualizing and interpreting new information with respect to established truths.

## 2. Material and methods

The analysis is based on a corpus of 200 RAABs published in the period 2000-2010 in four different disciplines − namely Applied Linguistics (AL), Economics (EC), Law (LA) and Medicine (ME) − taken from CADIS (Corpus of Academic Discourse, compiled at the University of Bergamo, cf. Gotti 2006, 2012). The corpus totals 42,730 running words of which 9,865 are found in 50 AL abstracts (ALABs), 7,063 in 50

EC abstracs (ECABs), 10,001 in 50 LA abstracts (LAABs), and 15,801 in 50 ME abstracts (MEABs).

The analysis is meant to identify and examine those linguistic items which are used to portray the construction of disciplinary knowledge, to introduce concepts and methods, and to represent evidence and its interpretation. Such expressions may variably thematize knowledge-making as an activity of observation (through "research acts"), of interpretation (through "cognitive acts"), or of argumentation (through "discursive acts") (cf. Thompson and Ye 1991; Thomas and Hawes 1994; Hyland 2002; Hyland and Tse 2005). The group of thematizing strategies that will be investigated here is constituted by:

- verbs of research (i.e. *show, demonstrate, observe, prove,* etc.), cognition (i.e. *hypothesize, think, assume, suggest, imply,* etc.) and discourse (i.e. *argue, claim, maintain, conclude, contend*, etc.);
- verbal and phraseological expressions identifying observation acts (i.e. *take a close look, tease out*, etc.), cognitive acts (i.e. *it is possible that, the belief is that*, etc.) or acts of discourse (i.e. *trace a line, move on to, the argument is that,* etc.);
- lexical items denoting research (i.e. *test, evidence, result,* etc.), cognition (i.e. *belief, hypothesis, assumption,* etc.) or discourse (i.e. *conclusion, claim, explanation,* etc.).

Word searches were both automated and manual. The manual processing of the texts was required because of the need to verify the pragmatic function of the various instances within their co-text and disambiguating spurious cases. WordSmith Tools 5.0 (Scott 2007) was used for quantitative searches.

## 2.1. Handling ambiguous cases

The first step in the analysis was the identification of the various markers and, contextually, the disambiguation of dubious cases, so as to be able to classify all thematizing resources as either research, cognitive or discourse acts on the basis of the role they play within their co-text and collocational pattern. For the sake of exemplification, two of the most problematic cases and the relevant strategies for their disambiguation will be illustrated in the paragraphs which follow.

One such case is represented by the verb *suggest*, which is often classified among communication verbs (Hyland and Tse 2005; Bondi

2010). In fact, its pragmatic function may vary considerably according to the co-text, as can be seen by observing the following examples:

**(1)**
This article **suggests** that the notions of code specificity and code-specific genre can be useful ones for theorizing the relationship between code and communicative practice in bilingual/multilingual settings. (ALAB12)

**(2)**
**It is suggested** that a transition is taking place towards new modes of organising transnational corporations' innovative activities. (ECAB50)

**(3)**
In addition, our results **suggest** that investment in financial assets is positively associated with returns to human capital investment. (ECAB05)

In all cases the semantics of the verb presupposes a certain degree of expertise (cf. Bondi 2010) on the part of both the writer and the reader for them to be able to make inferences on the basis of given data. However, despite having such a marked interactional character, which implies meaning negotiability, the function of the verb is different in each of the three examples. Indeed, when combined with volitional subjects or their metonymical representation (i.e. *the article suggests* for *the author suggests*, as in (1)), or with passive and impersonal forms, as in (2), where agency is concealed but not ruled out, *suggest* has a clear communicative function, working as a mitigated form corresponding to the expression *the author claims*. Such occurrences will be considered here as discourse acts. On the other hand, when paired up with non-volitional subjects denoting research components (i.e. *results, data, case study, evidence, outcomes,* etc.) as in (3), *suggest* functions as an interpretation act, in that it points to one among the possible understandings of the scientific value of a given piece of evidence. For this reason similar instances are accounted for here as acts of cognition.

Another ambivalent item is the verb *find*, both from a semantic and pragmatic perspective, in that it can be found in drastically different contexts as the excerpts below show:

**(4)**
I **find** that two instruments are used to deter illegal extraction: policing efforts and purposeful 'overexploitation' of the resource. (ECAB17)

**(5)**
We **found** that nitrotyrosine levels in idiopathic dilated cardiomyopathic (DCM) hearts were almost double those of control hearts in age-matched groups. (MEAB20)

As can easily be seen, in instances like (4) the verb is used with the meaning of *think* or *believe* and will be here counted among cognitive acts, whereas in (5) *find* has the meaning of *discover*, thus being distinctively a research act. The same process of disambiguation has been applied to a variety of other verbs such as *imply, indicate, mean, point to, presuppose*, *study*, *evaluate, measure,* etc., so as to provide a sound basis for our analysis and discussion.

## 3. Results

### 3.1. Conceptualizing knowledge-making

As mentioned above, when conceptualizing and synthesizing disciplinary knowledge and knowledge-making, scholars may choose to rhetorically frame both phenomena according to three different epistemological frameworks, each characterized by the presence and predominance of either "research acts", "cognition acts" and "discourse acts".

Research acts are those linguistic resources which point to experimental activities, representing the process of knowledge-making as though resulting from data observation. Such a function is typically carried out by verbs both denoting or presupposing a human and volitional agent (i.e. *I/we/the author/the study + analyse(s), investigate(s), show(s), demonstrate(s), test(s), find(s)*, *explore(s), examine(s), study(-ies),* etc.) and phenomenological evidence (i.e. *data/evidence/results + indicate(s), show(s), reveal(s), prove(s),* etc.), as can be seen in the examples below:

**(6)**
We **analyse** the pricing and informational efficiency of the Italian market for options written on the most important stock index, the MIB30. (ECAB38)

**(7)**
Sequential analyses **show** how code-switching works to escalate social opposition, often to the peak of an argument, resulting in subsequent backdown or full termination of the dispute. (ALAB49)

**(8)**
The results **indicate** that students engaged in a high degree of interactivity as well as all types of social and cognitive presence. These findings **indicate** that students not only progressed in their cognitive understanding of the pedagogical topics […]. (ALAB02)

In the present study, for a comprehensive understanding of the function played by such verbs, the full range of their textual realizations − in terms of tense, mood and modality − has been considered and counted, that is, those instances where knowledge-making practices are expressed in the past tense, in the passive or impersonal form, or through hedging formulations, since the choice of such resources − rather than the assertive active-voice present tense − is likely to modify denotation, as is exemplified in the excerpts below:

**(9)**
Over half of feedback moves led to immediate repair. Negotiation moves **proved** more effective at leading to immediate repair of errors than did recasts. (ALAB30)

**(10)**
Patients **were screened** with TCD, and if MES were detected, they **were randomized** to clopidogrel and aspirin or aspirin monotherapy. (MEAB23)

**(11)**
Strikingly, **it is shown** that many tokens are realized with highly subtle cues, involving only minimal drops along the f0 and amplitude dimensions. (ALAB13)

**(12)**
The technique permits general utility functions that may or may not be time-separable. […] Divergent borrowing and lending rates **can be handled**, as can stochastic labour income risks. (ECAB04)

As we can see, the past tense in (9) is used to confer definiteness to the claim in that it points to an experiment which is rhetorically represented as having been concluded and whose results are thus to be taken as established truths. The passives in (10) and (11) are meant for depersonalization purposes, by emphasizing objectivity both in describing the research and evaluating its results (cf. Shaw 1992). Finally, the use of hedges in (12) is meant to mitigate assertiveness and infer caution.

Research acts can also be introduced by phrasal expressions or autonomous lexical items (i.e. *test, evidence, results, findings,* etc.), as in the cases below:

**(13)**
**We combined molecular and biochemical approaches to identify** a functional variant of the CYP4A11 20-HETE synthase and determine itsassociation with hypertensive status in 2 independent human populations. (MEAB17)

**(14)**
An explanation based on **phonological knowledge** is posited instead. (ALAB07)

**(15)**
Finally **there is some evidence** that required rates of return were declining during this period. (ECAB02)

As can be seen, formulations like those emphasised in the excerpts above are or contain keywords whereby knowledge is associated to or represented as observation, investigation or experimentation, thus as a research activity.

The frequency and the distribution of research acts across our corpus are expressed both in absolute and normalized terms (calculated per 10,000 words) in Table 1. For a detailed discussion, results are listed according to their occurrence in the active vs. passive voice and the present vs. past tense.

| | ACTIVE | | PASSIVE | | |
|---|---|---|---|---|---|
| | PRESENT | PAST | PRESENT | PAST | TOTAL |
| ALAB | 92 | 26 | 17 | 9 | 144 |
| *normalized* | *93.1* | *26.2* | *17.5* | *8.7* | *145.5* |
| ECAB | 104 | - | 11 | - | 115 |
| *normalized* | *146.5* | - | *15.8* | - | *162.3* |
| LAAB | 69 | 3 | 8 | - | 80 |
| *normalized* | *68.8* | *2.7* | *8.2* | - | *79.7* |
| MEAB | 79 | 59 | 6 | 59 | 203 |
| *normalized* | *49.9* | *37.6* | *3.9* | *37.5* | *128.9* |

**Table 1. Occurrence of research acts.**

A general observation about the data in Table 1 is that in RAABs knowledge is presented as research more consistently in those disciplines which are of a distinctive hypothesis-testing nature, namely AL (145.5 per

10,000 words), EC (162.3) and ME (128.9). Among such disciplines, those which are typically experimental and corpus- or case-based, namely AL and ME, resort more frequently to the past tense, as if truth resulted from a perfected experimentation, thereby rhetorically limiting the risk of subjective interpretation. In the specific case of ME, objectivity is emphasised by a quite consistent use of passives, as if data and their value were self-evident and independent from an evaluating agent. In ECABs knowledge textualization, mainly in the present tense and active voice, is much more researcher-responsible: writers describe how they carried out their research and what they have found. A completely different trend is instead found in LAABs where research acts are quite limited and usually framed in a personal perspective. Cognition acts are those expressions used to represent mental processes and interpretive activities, and are typically realized through cognitive verbs (i.e. *think, believe, assume, hypothesize, expect, know*, etc.) either associated to a volitional agent (mostly *I, we, the author(s),* etc.) or, as we have seen above, terms referring to research components (i.e. *data, evidence, analysis,* etc.) which presuppose a preferred or expected interpretation (i.e. *suggest, imply,* etc.). As with research acts, our searches have considered occurrences in both the present and the past tense, active and passive voice, and also through mitigating resources, as can be seen in the examples below:

**(16)**
We **hypothesized** that in patients with germline mutations, BMPR2 might behave as a classic tumor suppressor gene, with somatic loss of the wild-type allele contributing to disease progression. (MEAB09)

**(17)**
Although interventions combining patient education and post-discharge management **have** demonstrated benefits in patients with chronic heart failure, the benefit attributable to patient education alone **is not known**. (MEAB14)

**(18)**
The cross-dialect propensity to have high coda glottalization rates before sonorant consonants **can then be understood** to arise as a phonetically natural consequence of normal coarticulation processes. (ALAB18)

Other cognitive acts are expressed by thematizing phrasal constructions (such as *it is uncertain whether, it is possible that, it is probable that,* etc.) which point to possible interpretations of a state of affairs rather than assessments of its validity in terms of certainty, as can be observed in the following examples:

**(19)**
Although CMs can be derived from hESCs ex vivo, **it remains uncertain** whether a functional syncytium can be formed between donor and recipient cells after engraftment. (MEAB16)

**(20)**
It also produces a new result: **it is possible that** an increase in the rate of change of labour productivity may not lead to an increase in the rate of change of employment. (ECAB37)

**(21)**
Primary percutaneous coronary intervention (PCI) has recently become the treatment of choice for AMI, but **it is still unknown** whether it has favorable effects on these prognostic variables. (MEAB30)

Finally, knowledge-making can be lexicalized as cognition through terms denoting or associated to the mental operation of conceptualization, comparison, or anticipation (i.e. *idea, belief, presupposition, intuition, hypothesis*, etc.), whereby states of affairs are treated as mental projections rather than measured as empirical evidence, as evidenced in the texts below:

**(22)**
**The intuition is that** when firms compete in licensing fees, resultant low licensing fees discourage firms from licensing to outside firms. (LAAB65)

**(23)**
Discussed in particular is the **hypothesis** that manifestations of relational practice differ in distinct communities of practice, and the validity of the equation of relational practice with "feminized" discourse is questioned. (ALAB17)

**(24)**
The study is based on the **belief** that investigations of the natural patterns of new technology use by ethnic communities will help us understand how technology could be involved in initiatives aimed at increasing the levels of language transmission and maintenance. (ALAB11)

Occurrences of cognition acts are listed in Table 2. According to the evidence collected in Table 2, the stage of data interpretation, which is fundamental in all knowledge-making contexts, is rhetorically represented as such mainly in ALABs, whereas the other disciplines make a more limited use of cognitive strategies, possibly because of their implicit subjective (thus face-threatening) character. However, the relevance of

such resources will acquire some significance when considered comparatively with the other thematizing formulations in section 4 below.

| | ACTIVE | | PASSIVE | | |
|---|---|---|---|---|---|
| | PRESENT | PAST | PRESENT | PAST | TOTAL |
| ALAB | 26 | 9 | 6 | | 41 |
| *normalized* | *26.6* | *8.7* | *5.8* | - | *40.7* |
| ECAB | 11 | 3 | | | 14 |
| *normalized* | *15.8* | *3.9* | - | - | *19.7* |
| LAAB | 19 | 3 | 3 | | 25 |
| *normalized* | *19.2* | *2.7* | *2.7* | - | *24.6* |
| MEAB | 28 | 6 | | | 34 |
| *normalized* | *17.8* | *3.9* | - | - | *21.7* |

**Table 2. Occurrence of cognition acts.**

Discourse acts are those resources by which knowledge is represented as originating from clear and coherent discoursal or reporting practices, as an argumentative construct or an explanatory activity (through verbs such as *acknowledge, argue, assert, claim, state, conclude, report, contend, ask, answer,* etc.) as can be seen in the following examples:

**(25)**
In **answering** the question, 'what distinguishes heterodoxy from the orthodoxy?', the **author argues** that matters of ontology are central. (ECAB19)

**(26)**
We also **describe** a FPAH patient carrying biallelic constitutional missense mutations of BMPR2 who manifested disease at a stage and manner similar to heterozygous patients. (MEAB09)

**(27)**
We **conclude** that electrically active, hESC-derived CMs are capable of actively pacing quiescent, recipient, ventricular CMs in vitro and ventricular myocardium in vivo. (MEAB16)

More frequently than with other knowledge-making acts, those concerning discourse are realized through quite articulated phraseological forms which are meant to illustrate (also in metadiscursive terms) the stages, mechanisms or dynamics of the coherent organisation, exposition and explanation of informative material, as is evident in the following excerpts:

**(28)**
The **work outlines a model** of constitutional adjudication in cases of conflict between these 'higher' forms of obligations in accordance with a deliberative understanding of the nature of the system of international law. (LAAB14)

**(29)**
The analysis is **formulated within a more general perspective** that also considers vertical structures. (ECAB07)

**(30)**
Having **recalled, in Section 2,** Europe's marginal role in the foundation of the UN at the end of World War II, and the fragmented existence of Europe in the Organization in the long period of the Cold War (Section 3), **the article turns to its central subject** – Europe's compliance with the rules of the UN Charter. (LAAB04)

Finally, discursive acts can also be performed by lexical items which activate frames of reference directly connected to the activity of arguing, explaining, illustrating or organizing information (i.e. *argument, conclusion, claim*, etc.), as seen in the extracts below:

**(31)**
However, the **argument** here concludes with an attempt to introduce a reflective critical dimension to sociological realism not through normative and analytical positivism, but through phenomenological analysis. (LAAB03)

**(32)**
To **illustrate my arguments (D)**, I examine (R) in depth 2 research programs developed by my colleagues and me over the last decade: research on extraversion as a psychological variable investigated (R) within the tradition of individual differences in SLA, and research on the expression of emotion in the L2. (ALAB09)

**(33)**
The central **claim** is that, from the perspective of an agency subject to judicial review, textual plausibility and procedural formality function as strategic substitutes. (LAAB55)

The results of our word searches are listed in Table 3. Table 3 indicates that discourse acts in the active voice are a strategic resource by which to thematize knowledge markedly in LAABs (128.8 per 10,000 words) and also in ECABs (71.3). The type of evidence examined in the two domains (i.e. human behaviour in the case of LA, and business-based or financial

phenomena highly dependent on human behaviour in EC) is very much culture- and situation-bound, therefore their scientific acceptability hinges on the coherence and cohesion of their rhetorical presentation. ALABs also resort extensively to discursive thematization but, unlike LAABs and ECABs, here the use of the passive is statistically relevant (representing almost one third of all occurrences): being a typical corpus-based discipline, evidence is presented as being self-explanatory rather than made rhetorically compelling by an expert arguer. The evidence-based character of ME is also retrievable from the data in the table, since MEABs minimize the use of discourse acts, notably balancing the present tense active voice (15.9) with half of the occurrences being either in the past active (1.9) or in the past passive (5.9), thus connoting knowledge as unbiased or contaminated very little by human agency.

| | ACTIVE | | PASSIVE | | |
|---|---|---|---|---|---|
| | PRESENT | PAST | PRESENT | PAST | TOTAL |
| ALAB | 69 | 6 | 20 | | 95 |
| *normalized* | *69.8* | *5.8* | *20.4* | - | *96.0* |
| ECAB | 50 | - | 6 | | 56 |
| *normalized* | *71.3* | - | *7.9* | - | *79.2* |
| LAAB | 129 | - | 3 | 3 | 135 |
| *normalized* | *128.8* | - | *2.7* | *2.7* | *134.2* |
| MEAB | 25 | 3 | - | 9 | 37 |
| *normalized* | *15.9* | *1.9* | - | *5.9* | *23.7* |

**Tables 3. Occurrence of discourse acts.**

## 4. Discussion

The present section opens by investigating affiliating strategies, that is to say, those resources exploited by writers with the purpose of placing their scientific activity within a recognizable scholarly tradition and, in particular, configuring the activity as an advancement in line with the processes and practices of knowledge-production typical of a given discipline. In RAABs the function of situating the research is generally carried out by the introductory section, that is the background move, which is usually fairly easily distinguishable from the rest of the text (it generally being the opening paragraph of the RAAB), as in the following cases:

**(34)**
The advantages or disadvantages of interjurisdictional competition are a hotly **discussed** topic in institutional economics. While the academic discussion is often concerned with the properties of an institutional meta-frame […], in this commentary a closer look is given to the "inner" constitution of jurisdictions […]. (LAAB 64)

**(35)**
Background. **It has been shown** that thrombin injection is a safe and effective technique for the treatment of iatrogenic femoral pseudoaneurysm. (MEAB47)

However, even when the introductory section is missing (not only when it is not metadiscursively labelled), informative elements pertaining to the background are scattered throughout the main body of the RAAB with the aim of introducing the informative scenario against which the new knowledge-constructing activity acquires meaning and relevance, representing or alluding to recognizable methodologies, established ideas and existing interpretations, as the following excerpts show:

**(36)**
These results can be contrasted with existing group mean unit root and cointegration tests that **indicate** sustainability for the group as a whole. (ECAB16)

**(37)**
While **previous research has assumed** that such repairs are vowel epenthesis, a detailed acoustic analysis indicates that inserted schwas are significantly different than lexical schwas. (ALAB07)

**(38)**
Whether **contemporary interventional treatment strategies** have improved outcomes for women compared with men **is unknown**. (MEAB02)

As we can see, whether it appears at the beginning of the text, metadiscursively labelled, or is dispersed throughout the whole body of the RAAB, background material can always be distinguished from the main and primary ideational content of the RAAB, that is, the core of the abstracted RA, the textual *raison d'être* of RAABs as a genre.

In order to be able to measure continuity and variation in thematizing strategies when presenting background and main information, in Tables 4 and 5, frequencies of knowledge-making acts are organised (respectively in absolute and normalized figures) according to their appearance in either

section (labelled as "background" and "main" in the table) and, for a more in-depth analysis, such occurrences have also been distinguished according to their realization in the active or passive voice, on the basis of which it will be possible to measure the degree of personal attribution on the one hand, and the self-evidence and well-established-ness of a knowledge claim on the other.

| | RESEARCH | | | COGNITION | | | DISCOURSE | | |
|---|---|---|---|---|---|---|---|---|---|
| | active | passive | *TOTAL* | active | passive | *TOTAL* | active | passive | *TOTAL* |
| ALAB | | | | | | | | | |
| *background* | 29 | 2.9 | *32* | 9 | 3 | *12* | 11 | 3 | *14* |
| *main* | 89 | 23 | *112* | 23 | 6 | *29* | 63 | 20 | *83* |
| *TOTAL* | | | *144* | | | *41* | | | *97* |
| ECAB | | | | | | | | | |
| *background* | 11 | - | *11* | 3 | - | *3* | 8 | - | *8* |
| *main* | 92 | 11 | *103* | 11 | - | *11* | 42 | 6 | *48* |
| *TOTAL* | | | *114* | | | *14* | | | *56* |
| LAAB | | | | | | | | | |
| *background* | 6 | - | *6* | 6 | - | *6* | 19 | 3 | *22* |
| *main* | 66 | 8 | *74* | 16 | 3 | *19* | 110 | 3 | *113* |
| *TOTAL* | | | *80* | | | *25* | | | *135* |
| MEAB | | | | | | | | | |
| *background* | 16 | 6 | *22* | 13 | - | *13* | - | - | - |
| *main* | 125 | 53 | *178* | 22 | - | *22* | 25 | 13 | *38* |
| *Total* | | | *200* | | | *35* | | | *38* |

**Table 4. Distribution of thematizing acts in background and main RAAB sections, expressed in absolute figures.**

As can easily be observed, all disciplines considered here tend with no exception to denote and connote the background of the knowledge-making activity, present existing claims and portray domain-specific practices, through the same thematizing resources then exploited in the main part of the RAAB. Or, from a different angle, we may say that the main part of the RAAB tends to replicate the same thematization used to set the scientific activity within some recognizable traditions and practices, thus creating a sort of epistemological continuum between existing, accepted and consolidated knowledge and new information. More specifically, we see that ALABs and ECABs introduce the domain as being characterized primarily by research activities (respectively 32.0 and 18.0 occurrences of research acts), while discourse (respectively 13.5 and 11.8) and cognitive practices (11.6 and 3.9) play a secondary instrumental role, and an equivalent distribution, though amplified in quantity, can be found in the main body of the RAABs of both disciplines (113.5 research, 29.1 cognition and 82.5 discourse acts in ALABs; 146.5 research, 15.8 cognition and 67.4 discourse acts in ECABs). In LAABs background knowledge is presented as a discursive construct (21.9) and the same epistemological framework is used to represent new claims (112.3), although with a marked increase in the use of research acts (from 5.5 to 74.2). Finally, MEABs refer to existing knowledge as a research-based construct (14.9) and frame new claims according to the same parameters (114.0), this time with a noticeable increase in the use of discursive acts (from 0 to 23.7).

These forms of continuity indicate, as has often been noted (cf. Swales 1990; Bhatia 1993; Duszak 1997; Hyland 2000; Gotti 2003) that writers, in order to corroborate the validity and trustworthiness of their contents and the pragmatic effectiveness of their textualization, tend to ascribe the activity being abstracted within a tradition and representation paradigm typical of a given discipline. In the light of such a claim, quantitative variation in the use of thematizing strategies in background and primary moves within RAABs of the same domain becomes particularly significant in that it is symptomatic of the way the new activity is to be understood and conceptualized with respect to established knowledge, and, more precisely, whether it refers to ongoing research, its interpretation or further argumentation in the case of new evidence. Each one of such choices presupposes varying degrees of confidence and caution when it comes to commenting upon, expanding or revising disciplinary accepted truths.

| | RESEARCH | | | COGNITION | | | DISCOURSE | | |
|---|---|---|---|---|---|---|---|---|---|
| | active | passive | *TOTAL* | active | passive | *TOTAL* | active | passive | *TOTAL* |
| ALAB | | | | | | | | | |
| *background* | 29.1 | 2.9 | *32.0* | 8.7 | 2.9 | *11.6* | 11.1 | 2.4 | *13.5* |
| *main* | 90.2 | 23.3 | *113.5* | 23.3 | 5.8 | *29.1* | 63.0 | 19.5 | *82.5* |
| *TOTAL* | | | *145.5* | | | *40.7* | | | *96.0* |
| ECAB | | | | | | | | | |
| *background* | 15.8 | - | *15.8* | 3.9 | - | *3.9* | 11.8 | - | *11.8* |
| *main* | 130.7 | 15.8 | *146.5* | 15.8 | - | *15.8* | 59.5 | 7.9 | *67.4* |
| *TOTAL* | | | *162.3* | | | *19.7* | | | *79.2* |
| LAAB | | | | | | | | | |
| *background* | 5.5 | - | 5.5 | 5.5 | - | 5.5 | 19.2 | 2.7 | *21.9* |
| *main* | 65.9 | 8.3 | *74.2* | 16.4 | 2.7 | *19.1* | 109.6 | 2.7 | *112.3* |
| *TOTAL* | | | *79.7* | | | *24.6* | | | *134.2* |
| MEAB | | | | | | | | | |
| *background* | 10.4 | 4.5 | *14.9* | 7.9 | - | *7.9* | - | - | - |
| *main* | 80.4 | 33.6 | *114.0* | 13.8 | - | *13.8* | 15.8 | 7.8 | *23.7* |
| *Total* | | | *128.9* | | | *21.7* | | | *23.7* |

**Table 5. Distribution of thematizing acts in background and main RAAB sections, expressed in normalized figures.**

In the case of ALABs, once the predominance of research acts in both background and main sections has been ascertained, by observing the total occurrences we note that the group of resources which increases proportionally and quite remarkably in thematizing new information is that of discourse verbs (from 13.5 to 82.5), whereas the increase in the number of observation verbs (from 32.0 to 113.5) and interpretation verbs (from 11.6 to 29.1), if noticeable, is statistically less relevant. This portrays the abstracted scientific activity as a discursive process which gathers data from observation, then organises, problematizes and explains them in ways intended to make them meaningful for the expert community. In other words, ALABs still inscribe knowledge-making within the research framework, in line with the corpus-based nature of the domain, but whereas existing knowledge is identified as the product of observation, new knowledge is lexicalized as a varyingly articulated argumentation of the observed evidence. This finds confirmation in the fact that, as we are aware, in AL RAs (the present chapter being a case in point) the Discussion section is fundamental and represents the key to the meaning of the whole text and, as a consequence, the associated RAABs emphasise this part of the scientific activity when abstracting its value. Another interesting piece of evidence is the increasing frequency in the use of passives in lexicalizing the main information, a strategy that is used rhetorically to conceal agency. The use of such a resource is likely intended for the purpose of reinforcing objectivity, and, more specifically, to compensate for, on the one hand, the "subjective" character represented by argumentation − which is primarily dependent on the writer's rhetorical ability rather than the empirical truth of the material observed − and, on the other, the inter-subjective and negotiative nature of cognition acts (here used much more frequently than in other disciplines) − and which appeal to a reader with the same cognitive competence as the writer's and presuppose the recognisability (i.e. the availability and accessibility) of the same logical mechanisms and models of reference allowing interpretation and understanding of given phenomena. Passives are thus used to rhetorically hide or circumscribe the face-threatening presence of an arguer by emphasising the self-evidence and well-establishedness of certain claims.

In ECABs the most widely-used strategy is represented by research acts, which are by far the preferred thematizing resources (more than double the number of discourse strategies, the second ranking type of knowledge-framing expressions found in ECABs). Unlike ALABs, where research and discourse thematizations had a complementary and mutually-clarifying function − and where, as we have seen above, discourse was

metadiscursively presented as being instrumental to the presentation and explanation of what was observed or yielded by research acts − here knowledge-making is lexicalized predominantly as research in both background and main sections. The increase in the use of cognition and discourse markers, while these are present, does not seem to be particularly significant, such markers having instead a peripheral function. In ECABs the main purpose in presenting new information seems to be that of replicating the same parameters used to identify current knowledge, as if such a repetition − that is, the filtering of new material through disciplinary accepted representation categories − was an appropriate means of corroborating the validity of new ideational content. Interestingly enough, the use of "objectivity" passives is considerably contained and corresponds to a high degree of personalization (i.e. first person singular, exclusive plural, expressions such as *the author*, etc.). In this case, linking the research and its validity to a given agent does not seem to be face-threatening since the writer is predominantly textualized as a researcher, that is someone who reports data yielded by observation and testing.

The case of LAABs is quite remarkable in that LA seems to be the sole academic discipline among those considered here where the scientific activity is systematically represented as a discursive construct, and, more specifically, as the argumentation of data gathered through observation. Within such an epistemological framework, observation plays a secondary role, in that it is meant to provide material allowing arguing, explaining, discussing and, ultimately, persuading. This predominance is found both in the background material and, more markedly so, when introducing new claims. This means that in LAABs the argumentative nature of the discipline is clearly established when introducing the background, the territory and the niche (21.9 occurrences of discourse acts) and, within this paradigm new information is thematized, made acceptable and relevant, as a sort of an expansion and amplification of such discursive practices (112.3), i.e. meant to further organise data through textual-discursive parameters such as premises, conclusions, warrants, claims, counterclaims, syllogisms, analogies, exemplifications, etc. What strikes us as being particularly interesting is that interpretation and research acts are found with a similar frequency in presenting established knowledge, whereas the latter become statistically more relevant in the main part of the LAABs. This dynamic distribution introduces a major difference: where new knowledge is thematized as argued on the basis of observation, established knowledge is portrayed as being more interpretation-based. Since interpretation, like all forms of cognitive activity, bears the mark of

subjectivity, this character is implicitly transferred onto current knowledge, which is therefore connoted as likely to be biased or contaminated by some interpretive flaws or inadequacies. The epistemologic "paradigm shift" between cognition and research found in the formulations of new information is possibly meant to emphasise its validity: new claims based on observation are rhetorically justified and motivated as an attempt to fill informative gaps due to a subjective reading of given phenomena. In LAABs the impression of objectivity within the argumentative framework is achieved in particular by statistically containing cognitive acts and exploiting research acts rather than resorting to morpho-syntactic resources like the use of passives or impersonal constructions.

MEABs display a drastically different trend. The main focus is upon establishing knowledge as an eminently research-based phenomenon. What is relevant here is that background information is portrayed, on the one hand, as the result of testing and observation − respectively through active (10.4 occurrences) and passive research acts (4.5) − and, on the other, as interpretation (7.9). The introduction of such a subjective perspectivization is a necessary rhetorical step in that it mitigates assertiveness and emphasises caution, which is a greatly appreciated value in a domain whose experimental nature calls for some indication of possible error (Thue Vold 2006, 246). On the other hand, it signals areas of possible knowledge gaps or uncertainty which are worth investigating further, that is, in other words, a niche to be occupied. As with the case of LAABs, the frequency of cognition acts is proportionally less marked in the main part of MEABs, and new information is instead presented more consistently as a matter of evidence rather than a form of interpretation of such evidence. This is meant to rhetorically contain the risk of bias or mitigate the possibility of error due to erroneous subjective evaluation, while at the same time implicitly relating such a possibility to the realistic problematicity of the experimental situation. Objectivity is also corroborated by the use of passives, which emphasise self-evidence and shared knowledge rather than agency. Finally, a noticeable trend is represented by the increase in discursive markers in the main part of MEAB texts (from 0 to 23.7): this is due to the fact that these abstracts tend to reflect as closely as possible the associated RA, replicating its structure and metadiscursively referring to its organisation and the sequencing of its parts, describing how the material is textualized and reporting what the main hypothesis for the case-study is, and what the result and the conclusion are (Swales and Feak 2009).

## 5. Conclusion

The present chapter has investigated those linguistic strategies most frequently exploited in RAABs by scholars in the various domains in order to represent ideational material. This is achieved by placing the ideational content of the RAAB within recognizable rhetorical frameworks which reflect the discipline's epistemological requirements as to how knowledge and the process of its creation are to be conceptualized and discussed so they may be effectively interpreted by the targeted community. From the evidence collected here, the predominance and consistency in the use of either *research*, *cognition* and *discourse acts* in RAABs of the same discipline seem to have a three-fold function: they are at the same time informative, "conformative" and dynamic. The use of a given thematizing resource has an informative function when instrumental in transparently and coherently representing ideational material and faithfully portraying the type of scientific activity that has been carried out. So, in ALABs, ECABs and MEABs − that is, in corpus- or case-based and experimental disciplines − evidence is primarily studied and secondarily explained and interpreted, whereas in LAABs the discussion of abstract principles is only applied secondarily to observed evidence and practical cases. Research, cognition and discourse acts have also a conformative purpose since, beyond their referential function, they are also meant to reflect (and corroborate) the epistemology at the basis of the various domains − i.e. how knowledge is conceived, produced and evaluated. So, typically academia-based disciplines, of a marked pedagogical nature and aimed at practical use in academic contexts − namely AL, EC, ME − tend to link knowledge to research by thematizing it as the result of investigation and analysis, whereas domains such as LA, which are less characteristically academia-centered and instead are primarily targeted to professional contexts, place less emphasis on research and interpretation, representing knowledge as a rhetorical and argumentation-driven construct, thematizing truths as the outcome of a discursive process. Finally, knowledge-making acts in RAABs have a dynamic function on the basis of their differentiated use in presenting background and main information, which indicates how new claims are to be interpreted with respect to given ones. Thus, in ALABs new material acquires relevance as a thoroughly argued exposition whereas established truths are markedly a matter of observation. In ECABs validation is acquired by representing new claims through the same thematizing strategies used to portray current knowledge, that is, as a research activity, in LAABs by linking discussion to new evidence-based material, and in MEABs by placing less emphasis on interpretive acts,

which are instead found in the background, and emphasising coherence through discourse acts.

## References

Berkenkotter, Carol and Thomas Huckin. 1995. *Genre knowledge in disciplinary communication: Cognition / culture / power*. Hillsdale: Lawrence Erlbaum Associates.

Bazerman, Charles. 1988. *Shaping written knowledge: The genre and activity of the experimental article in science*. Madison, WI: The University of Wisconsin Press.

Bhatia, Vijay 1993. *Analysing genre: Language use in professional settings*. London/New York: Longman

Biber, Douglas, Stig Johansson, Susan Leech, Susan Conrad, and Edward Finegan. 1999. *Longman grammar of spoken and written English*. Harlow: Longman.

Bondi, Marina. 1997. The rise of abstracts. Development of a genre in the discourse of economics. *Textus* 10(2): 395-418.

—. 2010. Abstract writing: The phraseology of self-representation. In *Linguistic interaction in/and specific discourses,* ed. Marta Conejero López, Micaela Muñoz Calvo and Beatriz Penas Ibáñez, 31-48.Valencia: Editorial Universitat Politècnica de València.

Bondi, Marina, and Silvia Cavalieri. 2012. The evolution of the abstract as a genre: 1988-2008. The case of applied linguistics. In *Genre change in the contemporary world. Short-term diachronic perspectives,* ed. Giuliana Garzone, Paola Catenaccio and Chiara Degano, 43-55. Bern: Peter Lang.

Cross, Cate, and Charles Oppenheim. 2006. A genre analysis of scientific abstracts. *Journal of Documentation* 62(4): 428-446.

Dos Santos, Mauro. 1996. The textual organization of research paper abstracts in applied linguistics. *Text* 16(4): 481-499.

Duszak, Anna. 1997. *Culture and styles of academic discourse.* Berlin / New York: Mouton de Gruyter.

Gotti, Maurizio. 2003. *Specialized discourse. Linguistic features and changing conventions*. Bern: Peter Lang.

—. 2006. Creating a corpus for the analysis of identity traits in English specialised discourse. *The European English Messenger* 15(2): 44-47.

—. 2012. *Academic identity traits. A corpus-based investigation.* Bern: Peter Lang.

Hyland, Ken. 2000. *Disciplinary discourses: Social interactions in academic writing*. London: Longman.

—. 2002. Activity and evaluation: Reporting practices in academic writing. In *Academic discourse*, ed. John Flowerdew, 115-130. London: Longman.

Hyland, Ken, and Polly Tse. 2005. Hooking the reader: A corpus study of evaluative *that* in abstracts. *English for Specific Purposes* 24(2): 123-139.

Lorés Sanz, Rosa 2004. On RA abstracts: From rhetorical structure to thematic organisation. *English for Specific Purposes* 23(3): 280-302.

Salanger-Meyer, Françoise. 1990. Discoursal flaws in medical English abstracts: A genre analysis per research- and text-type. *Text* 10(4): 365-384.

Scott, Mike. 2007. *WordSmith Tools version 5.0*. Oxford: Oxford University Press.

Shaw, Philip. 1992. Reasons for the correlation of voice, tense, and sentence function in reporting verbs. *Applied Linguistics* 13(3): 302-319

Swales, John. 1990. *Genre analysis. English in academic and research settings.* Cambridge: Cambridge University Press.

Swales, John, and Christine Feak. 2009. *Abstracts and the writing of abstracts.* Ann Arbor: The University of Michigan Press.

Thomas, Sarah, and Thomas P. Hawes. 1994. Reporting verbs in medical journal articles. *English for Specific Purposes* 13(2): 129-148.

Thompson, Geoff, and Yiyun Ye. 1991. Evaluation in the reporting verbs used in academic papers. *Applied Linguistics* 12(4): 365-382.

Thue Vold, Eva. 2006. The choice and use of epistemic modality markers in linguistics medical research articles. In *Academic discourse across disciplines,* ed. Ken Hyland and Marina Bondi, 225-249. Bern: Peter Lang.

# CHAPTER FIVE

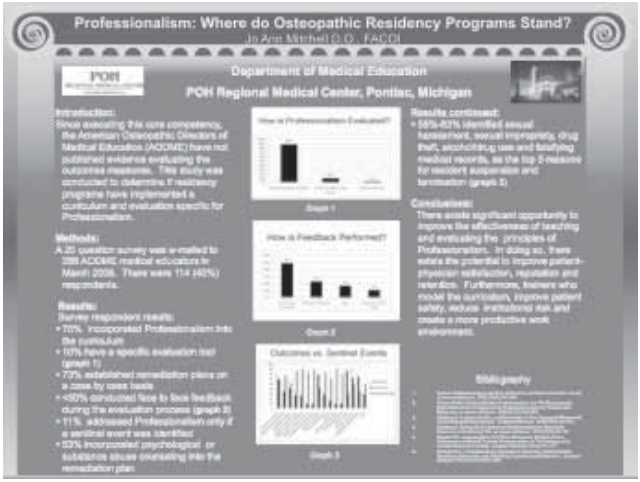
# "IF MSM ARE FREQUENT TESTERS THERE ARE MORE OPPORTUNITIES TO TEST THEM": CONDITIONALS IN MEDICAL POSTERS— A CORPUS-BASED APPROACH

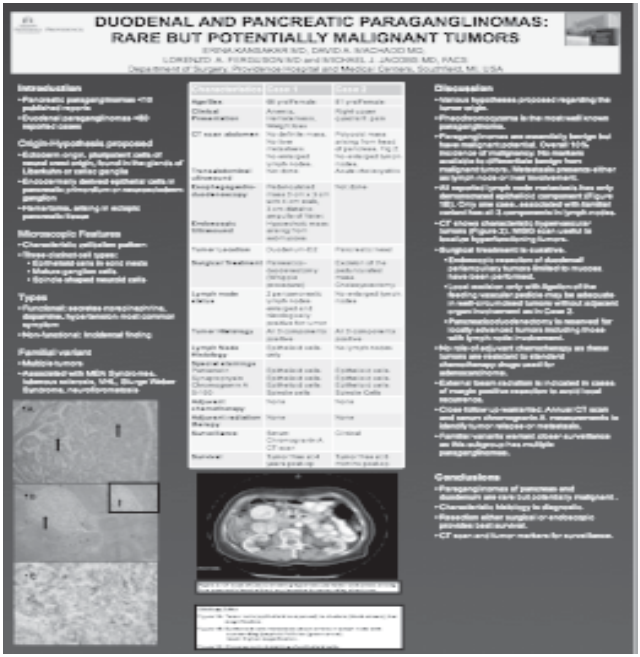
STEFANIA M. MACI
UNIVERSITÀ DI BERGAMO, ITALY

## 1. Introduction

Amongst the genres used by the medical community to spread its knowledge at conferences one of the most common is the poster. Poster sessions first made their appearance in the US in the 1970s, after being pioneered in Europe (Maugh 1974), and since then have acquired such a key role in scientific communication at medical conferences that they are valued more than oral presentations, because of the recognition that medical knowledge can better be absorbed from the former than the latter (Matthews and Matthews 1996, 97). Most poster guidelines define a poster as the visual display of research key points, usually presented at medical conferences during poster sessions by the researcher him/herself, or by one of the researchers if the poster is multi-authored. Indeed, as Dubois (1985a) claims, a poster is an alternative to reading a paper and can be a convenient way to communicate research results publicly at medical meetings. In such presentations, the presenter highlights key research points during interaction with the audience.

As we can see from Figures 1 and 2, posters can be produced in various sizes, contain varying numbers of tables and graphs, and, as indicated by poster guidelines, the presence of the abstract or the reference list is not compulsory. Since the purpose of a scientific poster is to share




**Figures 1 and 2. Poster samples taken from medical conferences.**

recent research findings with peers, and thus promote constructive discussion about the conception and results of a study (Demarteau et al. 2007, 91), communicating research findings via a poster involves much more than simply formatting issues (MacIntosh-Murray 2007). Indeed, while oral presentations contextualise a study against a background of relevant literature and the scientific protocols to be applied to an investigation, the findings of which will then be persuasively discussed by the author to validate the whole procedure, poster authors are required to focus more on a concrete message, evidence and titles, and to combine visual, oral and written elements to achieve scientific success (Matthews and Matthews 1996).

Poster sessions are particularly favoured by conference organising committees because the poster format permits the maximum number of presentations within a limited period of time (Pearce 1992, 1680). For instance, the *18th Congress of the European Academy of Dermatology and Venereology* held in Berlin, October 7-11, 2009, hosted over 1,200 congress papers in various parallel sections during the five days of the conference and well 1,563 poster presentations presented during the afternoon coffee-breaks (data retrieved from www.eadv.org). Poster presenters too seem to favour poster sessions rather than oral presentations, as suggested by Dr De Castro, Head of the Publishing Unit of the Istituto Superiore di Sanità, the National Health Institute of Italy (www.iss.it), whom I informally interviewed in 2009, and who told me that a poster presentation, which normally lasts no longer than ten minutes, involves face-to-face and relaxed interaction between poster author and poster viewer. As most posters present the preliminary results of embryo research, any scientific communicative exchange, feedback and new ideas the audience might offer can encourage development of the research project in the right direction. De Castro further informed me that poster presentations seem to be favoured by both novice and senior members of the academic community, for two other notable reasons: (a) prizes are generally offered for the best poster in terms of scientific novelty and originality, not only by medical foundations but also by pharmaceutical companies, which means having one's own research project funded; and (b) poster abstracts can be published in major specialised medical journals. This, of course, while complying with the 'publish or perish' (Wilson 1942, 197) *credo*, is seen as a convenient means to further academic careers. Indeed, the yardstick adopted to recognize scientific members' professionalism is based on communication of their own research through the Academia; such communication, which can be *shown* at conferences, should also be *printed*[1] in medical journals, by which means members of

the medical community acquire prestige and professional recognition (Dubois 1985b).

Medical literature about posters ranges from guidelines, e.g. advice on poster size, fonts, layout, colour and presentation, offered by medical associations and medical authors (Woosley 1989; Block 1996; Brooks-Brunn 1996; Keegan and Bannister 2003; Miracle 2003; Di Blasio 2004; Smith et al. 2004; AIFA 2005; Erren and Bourne 2007; Lorenzoni et al. 2007; Miller 2007; Willet et al. 2008; Ellerbee 2009; Hess et al. 2009; De Castro 2009; Purrington 2009; Rowe and Ilic 2009; Stoss 2010; American Heart Association 2011), to suggestions regarding the use of digital tools in order to improve the poster presentation (De Simone et al. 2001; Powell-Tuck et al. 2002; Zandifar et al. 2005; Dogan Bozdag 2008; Huang et al. 2008), and to the analysis of posters as a means for professional development (Miracle 2003), as well as for pedagogical reasons (van Naerssen 1984; Bracher et al. 1998; Hay and Thomas 1999).

The genre of medical posters has often been neglected from an applied linguistics perspective. An explanation may be found in Swales and Feak (2000) who note that posters do not have the same academic prestige as traditional papers. Their analysis, however, focuses on the pedagogical issue of helping their intended audience, novice writers in the linguistic field, to write (rather than to present) successful posters. Swales (2004, 21) further considers posters as a hybrid genre in which elements belonging to research papers, conference visuals or handouts are grouped together. In other words, he defines posters as a multimodal communicative event, with text, graphics, colour and (interactive) speech used to convey meaning. Yet, as claimed by MacIntosh-Murray (2007, 351-352), successful academic communication via posters means combining written andoral modes, which makes the genre of posters a very complex one, in that, on the one hand, academic writing is combined with editorial constraints and, on the other, academic discourse is mixed with informal interaction with an expected peer audience; such complexity is further increased by the fact that posters must be created in such a way as to stand alone and do the talking while showing medical research. MacIntosh-Murray's (2007) investigation draws on Dubois' studies (1985a, 1985b), which first examine the genre features of posters and evidence the popularisation function of medical posters in scientific communication being exploited to attract an *invisible college*[2] made up of (medical) professional members working on similar topics in order to create potential networks of research teams. Both scholars offer a description of the features of the poster as a genre based on the observation of poster

presentations at a departmental research day (MacIntosh-Murray 2007) and at a biomedical conference (Dubois 1985a, 1985b).

Given the minimal investigation of medical posters as a printed format from a linguistic perspective, it is the purpose of this study to examine the way in which medical discourse is textually organised in this genre. In particular, the aim is to examine how such linguistic traits as conditional constituents are realized in posters, since they play a key role in the elaboration of the reasoning typical of scientific discourse. Medicine is indeed an empirical science which exploits inductive reasoning, in that generalisation and theoretical abstraction derive from specific observations of certain phenomena. In scientific discourse, the construction of any inductive reasoning is supported by hedging devices, by means of which the author involves the audience in (written) interaction, with the purpose of promoting or circumscribing research claims, and by *if*-constituents, useful resources for introducing an hypothesis. This inductive reasoning is facilitated by the pattern *if P, Q2* establishing causal links or specifying "the precise conditions under which the research was carried out" (Carter-Thomas and Rowley-Jolivet 2008, 191). However, scientific English offers the possibility to develop this discourse with the so-called conditional '0' (*if P*-present, *Q2*-present), whereby the text points to the identification of timeless scientific laws (*if you heat ice, it melts*). Yet, in medicine, establishing such timeless laws is impossible, as research is always in progress and interpreting results is uncertain, also because conclusions and claims are grounded in statistics (Carter-Thomas and Rowley-Jolivet 2008, 192).

This contribution, based on a corpus of fifty medical posters presented at international conferences, published on-line and collected in the 2000s, seeks to investigate how *if*-conditionals are employed in posters within the discourse of medicine. In particular, the aim is to establish the most common/frequent pattern for conditional constituents in order to see how *if*-clauses are exploited in posters and if they can be substituted by more subtle patterns in this genre.

### 1.1. *If*-conditionals: Literature review and terminological remarks

To the best of my knowledge, few linguists have linked *if*-conditionals to the question of genre, apart from Facchinetti (2001), Ferguson (2001) and Carter-Thomas and Rowley-Jolivet (2008). Facchinetti (2001, 135-138), drawing on Quirk et al.'s (1985, 1089-1102) classification, qualifies conditional constructions in a corpus of modern legal texts, according to

their syntax ('first', 'second' and 'third' type conditionals), semantics (open and closed conditions, respectively, when the question of the fulfilment of the condition is left unresolved or not fulfilled at all) and discourse function ('content domain', i.e. whenever they express a causal relation between the states of affairs posed in the protasis or apodosis, e.g. *If tomorrow it rains, I will stay at home*; 'epistemic domain', according to which knowing about the truth expressed in the protasis is sufficient to know about the truth expressed in the apodosis, e.g. *if he's divorced, he's been married*; and 'speech-act domain', which considers conditionals as polite markers, as in *if I may say, I do not agree with him*).[3]

Ferguson (2001), on the other hand, specifically investigates conditional structures exploited in medical discourse. In his study, he found that conditional structures can be constructed without the protasis, as in *help me and I'll buy you a beer*, or in *in good weather, we go to the beach,* or as indicated by Ferguson (2001, 62): *Tell me the answer and I'll buy you a beer*; *With good weather, the roses will be out by June*. Furthermore, Ferguson highlights the fact that, in some cases, the secondary clause can be introduced by many conditional subordinators other than *if*, such as *unless*, *providing, assuming, supposing, as long as*, etc.[4] The findings of his investigation reveal that event conditionals are favoured in medical research articles, with emphasis on the use of epistemically constructed *if*-clauses in the *past* + *past* verbal pattern, whereas in medical editorials "there is the frequent use of conditional protases to qualify the scope of recommendations, to modulate predictions or prognostications, and to present cautious generalisations – all resulting in a very high proportion of modalised apodoses" (Ferguson 2001, 80).

Carter-Thomas and Rowley-Jolivet (2008, 193) identify the following macro-functions covered by *if*: the 'factuals category', which seems typical of medical discourse, establishing "facticity", found operationally when *if* can be substituted by *whenever*; the 'refocusing category', where hypothetical conditionals are exploited argumentatively in order to speculate about the observed scientific world, making space for different counterclaiming viewpoints; operationally, this includes also hedging devices, recommendations and concessive structures; 'discourse management category' is when the author metadiscursively guides the reader through the text and can be topic-marking or topic-shifting (these types of *if*-clauses are non-assertive: since they are polite directives, they do not concur in FTAs formation).[5]

These studies also take into consideration the position of the protasis with respect to the apodosis. The 'fronting' position, by means of which the protasis comes before the apodosis, is taken for granted (cf. for

instance, Quirk et al. 1985), and this indirectly supports the fact that inversion of protasis and apodosis ('delaying') is possible without there being any consequences. Although the conditional sentence (a) *If it rains, I will not go out* can be inverted to form (b) *I will not go out, if it rains*, the same does not seem to be possible in a conditional sentence like (c) *Fix it and I'll buy you a beer*, as the reversed form (d) **I'll buy you a beer and fix it* does not seem to make any sense.

However, the researchers explain when apodosis-protasis inversion is not possible, but fail to explain whether the inversion, if it is possible, differs semantically from the 'normal' pattern protasis-apodosis. According to Saeed (2004, 247), the fronting or delaying of the *if*-clause may depend on pragmatic considerations, such as whether the *if*-clause is presenting new information ('focus') or old information ('topic').

Drawing on Carter-Thomas and Rowley-Jolivet (2008), this study will consider the verbal pattern of the *if P*, *Q* conditional structure that occurs in medical posters, as well as the function such a structure has in this genre. In addition, delaying and/or fronting of the protasis will be checked in order to see whether the information is presented as 'focus'or 'topic'.

## 2. Methodological approach

In order to determine what types of posters should be collected, a background survey questionnaire was conducted online (http://freeonline surveys.com/rendersurvey.asp?sid=0o5c4btwxd6cfzk876238); also, thanks to the support of the editorial staff of the online journal *Va' Pensiero,* produced by Pensiero Scientifico Editore, an Italian scientific publishing house, this was publicized at http://www.pensiero.it/news/news.asp? IDNews=1110. Amongst the doctors who completed the questionnaire, four were available for interview. In addition, the Head of the Publication Unit of the ISS was contacted to gather information about the role of poster presentations within the academic medical community. In all cases, however, both survey respondents and interviewees emphasised the fact that posters are presented at conferences while a research project is still in progress and, preferably, if conference organizing committees are offering a poster award. As the Head of the ISS stated, apart from medical conference venues, the main place to find posters is on the Internet. She also underlined the fact that if the abstract of a poster is accepted for publication, this can further a medical career, so I decided to check what types of medical journals published the abstracts of posters. Since the first news item about medical posters was published in 1974 (Maugh 1974), I started my search from 1975 via MEDLINE (www.medline.cos.com/) and

PUBMED (www.ncbi.nlm. nih.gov/pubmed/), the most authoritative online databases containing citations and abstracts taken from health and medical journals, as well as *the Journal of Citation Reports (*http:// thomsonreuters.com/products_services/science/science_products/a-z/journal_ citation_reports/), which helped me to identify the journals with the highest impact factor (henceforth IF).[6] Amongst all the medical journals available, I found that the journal which began the publication of poster abstracts was the *American Journal of Epidemiology*. I therefore decided to concentrate on posters within the epidemiological field. I then searched on the Net for all available posters presented at congresses and published online by institutions and medical schools, as well as by online journals with an ISSN code and specialising in poster publication. I was thus able to collect 532 posters written in English and presented at scientific conferences between 2001 and 2011 from the following websites:

(a) Barts and The London NHS Trust, http://www.ihse.qmul.ac.uk/cme/bscmeded/poster/index.html;
(b) The 2011 International Conference on Meningitis, http://www.meningitis.org/posters;
(c) International Conference on Retroviruses and Opportunistic Infections, http://www.retroconference.org/;
(d) Istituto Superiore di Sanità, www.iss.it;
(e) eposternet, http://www.eposters.net/;
(f) F1000, http://f1000.com/posters.

As for (a), (b) and (c), the sites were selected as being representative of the Anglo-American medical tradition; as for (d), this website is representative of Italy; and (e) and (f) are online journals publishing posters only.

In order to create a framework for the analysis of conditional structures to be applied extensively to all posters, I decided first to conduct a pilot study. I therefore randomly selected a sample of fifty posters to be part of my pilot study, that is, about 10% of all the posters collected, forming a small corpus of 104,357 running words. All texts on the selected posters were then closely read in order to form an overall impression of their content and argumentative strategies. A quantitative analysis was carried out with WordSmith Tools (Scott 2004), by means of which I was able to check if and what elements introduced the protasis, i.e. *if*, *unless*, *as/so long as*, *assuming*, *except*, *given*, *in case*, *in the event that*, *just so*, *on condition that*, *provided/providing that, save*, *suppose/supposing* (cf. Quirk et al. 1985; Facchinetti 2001), *whenever*, and *whether*, the findings

of which were computed according to the standardised type-token ratio (STTR) per 10,000 words.[7] In addition, I also checked whether any modal expressions occurred in order to detect whether a conditional structure was constructed with or without a protasis. The quantitative analysis was later followed by a qualitative interpretation of the findings of this research on the basis of a previous descriptive and quantitative analysis, since the former implied several subtle distinctions, which will be discussed later.

## 3. Results

The first steps of my investigation has moved from Quirk et al. (1985) in order to see what type of elements introduce the protasis (Table 1).

| Elements introducing protasis | Hits (Standardised Type-Token Ratio*10,000) | | |
|---|---|---|---|
| | POSTER *104,357 tokens* | MEDICAL REFERENCE CORPUS *401,096 tokens* | BNC *454,188 tokens* |
| If | 32 (3.06) | 209 (5.21) | 749 (16.49) |
| Whether | 18 (1.72) | 123 (3.06) | 255 (5.61) |
| Given | 3 (0.28) | 30 (0.74) | 54 (1.18) |
| As/so long as | 1 (0.09) / 0 | 2 (0.04) / 1 (0.02) | 8 (0.17) / 0 |
| Except | 1 (0.09) | 3 (0.07) | 15 (0.33) |
| In case | 1 (0.09) | 3 (0.07) | 10 (0.22) |
| Assuming | 0 | 11 (0.027) | 16 (0.35) |
| In the event that | 0 | 0 | 0 |
| Just so | 0 | 0 | 0 |
| On condition that | 0 | 0 | 0 |
| Provided/providing that | 0/0 | 1 (0.02) / 0 | 5 (0.11) / 1 (0.02) |
| Save | 0 | 0 | 0 |
| Suppose/supposing | 0 / 0 | 0 / 0 | 6 (0.13) / 0 |
| Unless | 0 | 35 (0.87) | 57 (1.25) |
| Whenever | 0 | 5 (0.12) | 9 (0.19) |
| **TOTAL TYPES** | **57 (5.46)** | **423 (10.45)** | **1,185 (26.09)** |

**Table 1. Breakdown of elements introducing the protasis across (medical) genres.**

In order to see whether the trend found in my corpus was consistent with medical discourse in general, I carried out the same investigation, first on a corpus of medical texts of 401,096 running words comprising abstracts, research articles, clinical studies, and research letters published between 2000 and 2010 by *Jama*, *The Journal of the American Medical*

*Association* (*JAMA*) (IF 30), *The Lancet* (IF 32.49), *The New England Journal of Medicine* (IF 53.48)[8], which I will call the Medical Reference Corpus, and then with the BNC, in which my search was filtered to examine written academic publications published between 1960-1993, within the domain of applied science and medicine and targeting an adult audience, which allowed me to take into consideration a subcorpus of 454,188 words.

The three corpora align in terms of which elements introducing the protasis do not occur, namely *save, on condition that, in the event that, just so,* and *save*. Interestingly, *suppose/supposing* is an item absent both from the poster corpus and in the medical reference corpus.

It is worth noting the fact that, in the BNC corpus, STTR is a bit higher than in the Medical Reference Corpus. Presumably, the difference in frequency of the normalised figures in these two corpora lies in the fact that while the Medical Reference Corpus comprises just the macro-genre of the research articles (i.e. research/original articles, research letters; seminars; clinical cases – all of them structured according to the well-known 'Introduction/Methods/Results/Discussion' (IMRD)[9], the BNC filtered subcorpus comprises not only research articles in all their subgenres, but also editorials, comments, reviews, announcements, and so on, the latter having a less structured discursive pattern which may favour argumentation expressed with strategies other than *if*-conditionals, such as rhetorical questions, or concessive and adversative sentences. Indeed, the type of conditionals present in the Medical Reference Corpus is in line with both Ferguson (2001) and Carter-Thomas and Rowley-Jolivet (2008), who claim that *if*-clauses mainly occur in the 'Methods' section whenever they construct a cause-effect relationship (what Ferguson calls the 'event conditional' and Carter-Thomas and Rowley-Jolivet call 'factual conditional'), as in (1):

**(1)**
If an analysis of variance yielded a significant F test score, post hoc means comparisons were used to examine differences between specific groups. (*JAMA*, 2002)

and in the 'Discussion' section whenever they are 'hypothetical' (Ferguson 2001), thus belonging to the 'focusing category' (Carter-Thomas and Rowley-Jolivet 2008), as revealed in (2):

**(2)**
However, new scientific breakthroughs will be of no practical value if they are not made available at well functioning points of care. (*The Lancet*, 2010)

As for the minimal presence of the elements introducing the protasis in posters when compared to the Medical Reference Corpus and the BNC, the reason seems to lie in posters' main synoptical function: in fact, as Swales (2004) claims, posters are a hybrid genre combining visual and textual elements with oral presentation skills. In other words, the printed text on posters apparently relies more on visual elements, such as tables, charts, and pictures, than on verbal features to convey the scientific message. Indeed, 100% of my corpus presents the results in visual elements, whereas the text is left to the introductory and conclusive sections (when present). In addition, given the compressed form of the poster, which must respect space constraints, the text seems characterized by a less discursive narrative pattern than either research articles or the texts included in the BNC corpus: bulleted sentences make the poster text appear more assertive than argumentative, the latter feature being assigned to the poster authors' presentation. As said before, the poster content is displayed as a "visual unit" (MacIntosh-Murray 2007), all on a single view plane. Indeed, most conferences issue poster presentation guidelines with spatial limitations of 1m x 2m (4-ft x 8-ft) poster area.[10] Because of these limitations, poster text needs to be as concise as possible, reporting only the most important facts and key points. In other words, what in a research article is described, explained and argued in several pages, in a poster must be expressed in few words and condensed in a very limited amount of space.

## 3.1. Analysing poster protasis: IF

The analysis shows that *if* is the element that most frequently introduces the protasis. Its distribution varies across poster sections.[11] Indeed, there is a 36.6% frequency of *if* clauses in the 'Methods' section, 23.6% in the 'Introduction', where authors are stating the objective of their research or indicating the purpose of study, and just 16.6% of *if* subordinates in each of the 'Results' and 'Discussion' sections, as Table 2 indicates:

| Abstract | | | | | |
|---|---|---|---|---|---|
| Methods | Results | Introduction | Methods | Results | Discussion |
| 3.3% | 3.3% | 23.6% | 36.6% | 16.6% | 16.6% |

**Table 2. Distribution of *if* across poster sections.**

Closer analysis of such a distribution reveals that, syntactically, the preferred pattern in the 'Introduction' is *if present + could*: the conditional

pragmatically belongs to the 'refocusing category' (Carter-Thomas and Rowley-Jolivet 2008), in that it makes room to introduce the authors' niche in the existing literature, which is always counterclaimed, a commonly shared reality already acknowledged in the literature, as examples (3) and (4) suggest:

**(3)**
If such hotspots exist, they could influence the location of recombination events, particularly under circumstances in which the effective replicating pool of HIV is limited. (P012)

**(4)**
The purpose of this project is to evaluate if oxidative stress, induced by HIV-1, is able to determine alterations of telomeric structure and of telomerase activity in a human astrocytoma cell line and if antioxidant compounds can effect on telomeres shortening and telomerase activity. (P227)

The type of *if* conditionals in the 'Methods' section always follows the grammatical pattern *if present* + *present* and belongs to the 'factuals category', that is to the category employed when 'facticity' (Carter-Thomas and Rowley-Jolivet 2008) is the main prerogative in discourse, where the protasis introduces facts dependent on either the selection of subjects relevant to the experiment or the type of chemical substance injected into the subjects taken into consideration:

**(5)**
- Infected if assigned vaccine or placebo (always infected)
- Infected if assigned vaccine but not if assigned placebo (harmed)
- Infected if assigned placebo but not if assigned vaccine (protected)
- Not infected if assigned vaccine or placebo (never infected). (P069)

The 'Result' section of the posters belonging to my pilot study presents conditionals expressed in the sequence *if present* + *present*, showing, again, factual conditional sentences, where the protasis introduced by *if* is the cause whose effect is explained in the apodosis (6):

**(6)**
For example, if the vaccine efficacy is VES = 0**:**8, five animals per group suffice to obtain a power of more than 95%. (P060)

Interestingly, in two cases, the protasis is presented in a very peculiar way. In one poster (P061) the 'Results' section containing the protasis that

is expressed as the caption to a figure rather than being discursively organised within the poster text (Figure 3):

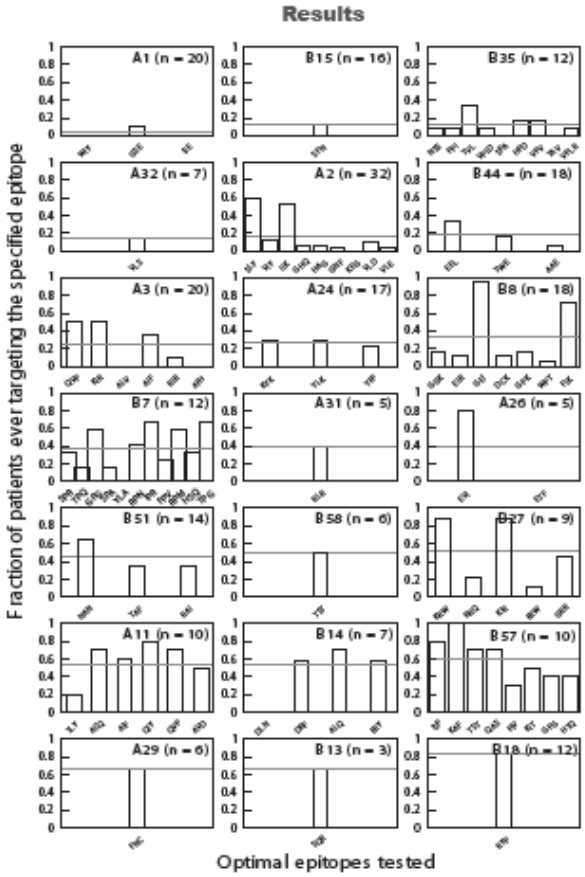

Results
A1 (n = 20)
B15 (n = 16)
B35 (n = 12)
A32 (n = 7)
A2 (n = 32)
B44 (n = 18)
A3 (n = 20)
A24 (n = 17)
B8 (n = 18)
B7 (n = 12)
A31 (n = 5)
A26 (n = 5)
B51 (n = 14)
B58 (n = 6)
B27 (n = 9)
A11 (n = 10)
B14 (n = 7)
B57 (n = 10)
A29 (n = 6)
B13 (n = 3)
B18 (n = 12)
Fraction of patients ever targeting the specified epitope
Optimal epitopes tested

Figure 3: Shown are the frequencies with which each individual optimal epitope was targeted in the patients expressing the restricting HLA. Patients were scored as responding to an epitope if they ever made a positive response during the Interruption Therapy Study. The panels of the different HLA alleles expressed in the cohort are ordered in terms of rising HLA-specific epitope targeting frequency (red line). The definition of the HLA-specific epitope targeting frequency is explained in figure 2.

**Figure 3.** ***Results*** **as caption.**

In another poster (P287), not only is the 'Results' section developed as the caption to a figure, but the protasis is also realised as the title of a figure whose apodosis is the figure itself (Figure 4):

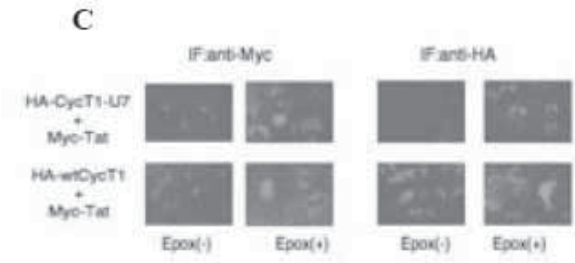



**Figure 4. Protasis-apodosis unconventional sequence.**

This seems to be a typical characteristic of posters, as the description of 'facticity' is meaningfully left to the visual representation. In other words, the language and discourse on posters are seen as metadiscursive tools necessary to position the visuals (tables, charts and/or pictures and drawings), which represent the results. In this way, the presence of the authors seems unnecessary, as the facts speak for themselves. This grants the utmost personal detachment and the highest scientific involvement (these pictures do not demand individual but rather scientific involvement, as in the case of Figure 2, they invite taking a microscopic look at the phenomenon described). The whole meaning carried by these visual elements is, however, not explained in any way but has to be interpreted by the reader him/herself, thanks to the relation created by the title-protasis. Thus the results, visually displayed, involve processes of supposition and inference in which the protasis functions as premise and the apodosis as conclusion (Ferguson 2001).

This very same view seems to be supported by the minimal presence of *if*-clauses in the 'Discussion' section (7):

**(7)**
This study is a useful starting point, but if LS EIA testing were extended to persons in NYC testing at private laboratories, estimates could be generalized to the population of all testers in NYC and better estimate population HIV incidence. (P055)

The typical pattern here is *if past* + *could/would* to reflect the argumentative speculation about the described world and belongs to the 'refocusing or hypothetical category' of modals (Carter-Thomas and

Rowley-Jolivet 2008). I would have expected a narrative discourse exploiting more argumentative strategies expressed by means of an *if P*, *Q2* structure, but apparently such a negligible presence of conditional constructions seems unnecessary because the authors' voice does not need to support facts which can speak for themselves.

The way in which information is structured within a conditional sentence is responsible for the order of the apodosis and protasis: we have 'fronting' when the protasis occurs before the apodosis, and 'delaying' when the protasis is after it. As Saeed (2004) and Carter-Thomas and Rowley-Jolivet (2008) suggest, 'fronting' occurs when the *if*-clause has a thematic role, since it expresses already known information, whereas 'delaying' is possible when the *if*-clause acquires a rhematic role and introduces new information. In my corpus, 'delaying' rather than 'fronting' is preferred, the former occurring in 75% cases, as can be seen in examples (8) and (9):

**(8)**
By assumption A3, everyone infected in the vaccine arm would have been infected if assigned placebo. Therefore all infected vaccine recipients are in the always infected principal stratum. (P069)

**(9)**
The purpose of this project is to evaluate if oxidative stress, induced by HIV-1, is able to determine alterations of telomeric structure and of telomerase activity in a human astrocytoma cell line and if antioxidant compounds can effect on telomeres shortening and telomerase activity. (P227)

Given the default practice of positioning the *if*-clause before the apodosis, postponing the protasis to a rhematic position means marking it and foregrounding the information it carries. In my corpus the standard and non-marked *if P*, *Q2* pattern occurs in eight cases (25%) and seems to be possible whenever predictivity is implied. Yet this does not seem the default position, as the remaining 75% of the *if*-clauses occurring in my corpus shows the postposition of the protasis, which is thus marked. As revealed by excerpts (8) and (9), the protases, which are in a rhematic position and therefore marked, indicate conditions influencing the results or hypotheses to be validated. In all instances of postponed *if*-clause, the marked protasis emphasises that the result cannot be predicted because it strongly depends on different factors. This thematic information acquires salience because inserted in a rhematic position. Therefore, the marked theme realised as the postponed protasis conveys new information which is thus emphasised and foregrounded as cognitively salient for the

audience. It seems, therefore, that in my corpus the default position of the *if P, Q2* is possible only when the results may be predicted, as it does not add any significative information; whereas the *Q*, *if (and only if) P* pattern seems to be the prescribed norm if the protasis carries the most relevant and unpredictable information in the text.

### 3.2. Analysing poster protasis: WHETHER

When the protasis is introduced by *whether* it always occupies a rhematic position and, in 50% of cases, is in the 'Introduction' section – also when the conditional clause is in the abstract:

**(10)**
This system will be used to determine whether hotspots for recombination exist in other regions of the genome and what characteristics they have in common. (P12)

**(11)**
In this study we used a transendothelial migration chamber (TEMC) with shear flow to test whether R5 gp120 affects the forward and/or retrograde transendothelial migration (TEM or retro-TEM) of T […]. (P486)

The pattern is always *infinitive + whether present* and normally introduces the aim of the study, thus belonging to the 'refocusing category' (Carter-Thomas and Rowley-Jolivet 2008).

The corpus presents 38.8% occurrences of *whether* to be in the 'Results' section (12):

**(12)**
We characterized the frequency, nature, and distribution of mutations in all sequenced clones, to determine the correlation between crossover events and mutations and whether certain segments of the analyzed sequence were more mutagenic than others. (P012)

**(13)**
Assumption A4: There is no selection bias (i.e., for infected placebo recipients, viral load is independent of whether or not they would have been infected had they received the vaccine). (P069)

The syntactical pattern in these cases is either *infinitive + whether present* or *present + whether would have*. In all cases, the conditional belongs to the 'refocusing' rather than the 'factuals category' (Carter-Thomas and Rowley-Jolivet 2008), as one would suppose, since the

*whether* clauses occur in the 'Results' section. Indeed, the poster authors, rather than reporting the results, are offering a justification for their operational scheme, which requires argumentative strategies.

In one case, the *whether* clause is present in the 'Conclusion' section of the poster, but this is possible only because the authors are summarizing the purpose of the study:

**(14)**
The aim of this study was to determine whether cell-free and cell-to-cell HIV transmission are mediated by different mechanisms […]. (P486)

In all cases but one, the *whether* clause is delayed, which means that the authors are foregrounding the protasis in order to emphasise that the results of the research being carried out strongly depend on the conditions put forward in the apodosis.

As to the distribution of *whether* across the IMRD pattern of posters, we can hypothesise that its higher presence in the 'Introduction' and 'Results' sections depends on the function the conditional clause has. Indeed, as said before, *whether* introduces 'refocusing' conditional sentences in the 'Introduction', where the aim of the study is indicated, and 'factual' conditional sentences in the 'Results' section, where the justification of the researchers is pointed out. The relative absence of *whether* in the 'Conclusion' section may be explained by the fact that here findings are summarized in a bulleted sentence style, where the style of the narration does not need to be argumentative but rather assertive. Indeed, as indicated in the introductory paragraphs of this contribution, posters must stand alone (McIntosh-Murray 2007) and facts must speak for themselves, whereas argumentation belongs to the oral presentation of the poster.

## 3.3. Analysing poster protasis: GIVEN, AS LONG AS / SO LONG AS, EXCEPT, IN CASE

Amongst the elements introducing the protasis listed by Quirk et al. (1985), my corpus presents *given*, *as long as / so long as*, *except* and *in case*:

**(15)**
Given that viruses depend on their host cells, surface protein distribution and activation state of target cells may affect the efficiency of infection. (P161)

**(16)**
Increased testing and condom use should be encouraged among all MSM. Data regarding risk behaviors among MSM are concerning, especially given the high HIV prevalence among this population. (P057)

**(17)**
As long as the rate of conversion to latency is > 1/105, latency is already established by the time virus-specific CD8+ T cell numbers begin to rise. (P064)

**(18)**
Panels C and D are the same as panels A and B, except cells were co-infected with independently generated parental stocks in a separate experiment. (P012)

**(19)**
Whereby VIF was supplied by the full length HIV-1 clone or in trans (in case of Δvif clones). (P404)

Table 3 offers an overview of their distribution across my corpus.

| Protasis | Introduction | Methods | Results | Conclusion |
|---|---|---|---|---|
| Given | 2 | | | 1 |
| As long as | | | | 1 |
| Except | | | 1 | |
| In case | | 1 | | |

**Table 3. Elements other than *if* introducing the protasis.**

Given their extreme low frequency in my corpus, no generalisation about their categorical use is possible. I may suppose that 'fronting' is possible when the protasis is predicting a possible result, and that 'delaying' seems to be required when the protasis is the necessary and sufficient condition for *Q* to be realised.

## 4. Conclusion

The use of conditionals documented in the presented analysis of medical posters tends to confirm Carter-Thomas and Rowley-Jolivet's (2008) claim that medical discourse employs a type of inductive reasoning path based on observation, which does not allow any deductive form of thinking, as findings are very frequently uncertain and inconclusive (2008, 192). In such discourse, the use of conditionals reflects the hard sciences

reasoning process: they can either convey 'facticity' or 'refocusing' (Carter-Thomas and Rowley-Jolivet 2008).

In the case of 'facticity', facts and results are reported according to the conditions they are associated with and are thus expressed in the 'Method' and 'Results' sections. Furthermore, factual *if* clauses can be predictive or can express the condition *sine qua non* a fact can occur. Their pragmatic role is indicated by the information ordering structuring: prediction seems to be realised with the fronting of protasis, whereas the *if and only if* condition appears to be constructed by means of delaying.

When conditionals argumentatively express the researchers' viewpoint, which hypothesizes the way in which the world they have observed is structured, *protasis-apodosis* clauses normally occur in the 'Discussion' section.

As Ferguson (2001) indicates, the frequent exploitation of *if* conditionals in the 'factual category', defined by the author as 'event conditionals', is the norm in medicine, since the "empirical observation of co-occurring events has a possibly more significant role than theoretical argumentation" (2001, 69). This may explain why the corpus does not present a significant presence of conditionals in the 'Discussion' section. Yet this does not seem to be justified by the cross-checking I carried out on the Medical Reference Corpus and the BNC where conditional constructions appear to be much more frequently employed (though in different measures) than on posters.

**Figure 5. Poster exhibition hall at the *2011 International Cancer Conference* held in Liverpool, 2011.**

However, if we consider the hybrid nature of printed posters, made up of two different modes, i.e. verbal text and visuals (tables, charts, pictures), in which space constraints force authors to prefer the visual element in order to attract the attention of the audience, given the

extremely high number of poster presentations hosted at conferences (cf. Fig. 5), it seems obvious that, in the discourse of medical posters, printed language loses its communicative and persuasive functions and acquires a merely metadiscursive role by means of which the audience is guided through the visuals to be interpreted.

In other words, if conditional constituents "involve processes of supposition and inference in which the protases function as premises and the apodoses as conclusions" (Ferguson 2001, 74), a genre such as the poster, in which the elaboration of conditional constructions is allowed in the form of hybrid structures and the premises in the protasis are formed by texts and the conclusion in the apodosis is realized with a magnified picture (as in Figure 4 above), it seems clear that language is unnecessary, as the representation of facts is reported by visual data which seem to speak for themselves and whose interpretation is left to the viewer of the poster, so the author's voice is just leading the way from one visual to another. The persuasive force of this strategy is greater than any other form of argumentation, since the apparent absence of the author's voice reinforces the sense of objectivity and impartiality which grants data authoritative status.

Such preliminary analysis certainly has some limitations, as it has involved the study of only fifty posters, and will surely need further investigation. Nevertheless, given the lack of applied linguistics research on posters, my study seems to offer unique insights into how *if*-conditionals are exploited in this neglected genre.

## Notes

[1] I thank Dr David M Svinarich Director of Patient Care Research and SJHS Research, and Development of Providence Hospital, Providence, U.S., for allowing me to reproduce two of the posters created by some 2010 Fellows and Residents and presented at medical conferences in 2009 (Fig.1) and 2010 (Fig. 2), respectively.

[2] In my corpus, for instance, 147 posters out of 532 have a reference list and 255 contain an abstract.

[3] The main characteristics of posters is that they are meant to stand alone, without the presence of the presenter. In other words, posters are supposed to *do* the talking (MacIntosh-Murray, 2007: 351-352) and *show* medical research. On the other hand, the abstract of a poster can be published in prestigious journals, which may enhance the poster authors' academic career.

[4] The concept of the *invisible college* has also been discussed in Kuhn (1962), Matthew and Matthews (2007), and Gross et al. (2002), who date its origins back to the seventeenth century.

[5] Facchinetti eventually carries out her analysis on legal texts by basing her investigation on syntactic and semantic conditionals which reveal that the most frequent conditional structures employed in legal texts are dependent on the contest and can be classified as either 'normative' or 'non-normative' conditionals.
[6] After summarizing conditional structures from both pedagogical and discourse function perspectives, Ferguson (2001) draws on Athanasiadou and Dirven's (1997) framework (quoted in Ferguson 2007), which distinguishes between course of event conditionals (as in *If he doesn't take the tablets, he feels dizzy*), hypothetical conditionals (enclosing 'first' and 'second' type conditionals, as in *If you come tomorrow, we will have a picnic*) and pragmatic conditionals (as in *If you had come, we would have had a picnic*).
[7] The authors identify some variations in the use of conditionals across genres, namely, medical research article (RA), conference presentation (CP), and editorial: they are, indeed, *factuals* in the RA, *refocusing* in editorials, and *discourse management* in the CP, which can be easily inferred, given the different purposes and functions these three genres have. In addition, syntactical variation in the conditional pattern can be seen according to the genre: in RA a common structure is *past* + *past*; in CP, given its oral quality, the exploitation of *present* + *present* sequences is customary; whereas in editorials, the use of the *if P, Q2* in all its variants is widespread , e.g. *if P, Q?*, *if P then Q*, *Q only if P* (Carter-Thomas and Rowley-Jolivet 2008, 203).
[8] The impact factor (I.F.) is a measure of the citation rate per article, this is used to indicate the importance of a journal in its field. See *ISI WEB of Knowledge* (available at: www.isiwebofknowledge.com/).
[9] Posters are not always formally divided into (*Abstract*) – *Introduction* – *Methods* – *Results* – *Discussion* sections. Sometimes they may have a section entitled *Background* instead of *Introduction*, or *Conclusions* rather than *Discussion* (cf. Maci 2009). Nevertheless, such an IMRD pattern is constantly recognizable when reading them. Therefore, I will refer to IMRD in posters for the sake of simplicity, rather than identifying single sections, which might result in confusion.
[10] In Corpus Linguistics, *tokens* are the running words of the corpus; *types* refer to each different kind of word in a corpus. TTR is the ratio between types and token. Since the resulting percentage may vary according to the length of the texts forming the corpus, the TTR is normally *standardized*; in other words, in order to be sure that the TTR represents fair results and percentages, the ratio is calculated for the first 10,000 running words, then calculated afresh for the next 10,000, and so on to the end of the text or corpus (see Hunston 2002, 17).
[11] Permission for corpus collection has been granted by all the above-quoted journals.

## References

AIFA - Ministero della Salute 2005. Il poster congressuale: Come prepararlo, illustrarlo e renderlo efficace. *Bollettino d'Informazione sui Farmaci* 5-6: 237-240.

American Heart Association 2011. *Poster Guidelines*. http://my.americanheart.org/idc/groups/heart-public/@wcm/@sop/@scon/documents/downloadable/ucm_431254.pdf

Block, Steven M. 1996. Do's and don'ts of poster presentations. *Biophysical Journal* 71: 3527-3529.

Bracher, Lee, Jane Cantrell, and Kay Wilkie. 1998. The process of poster presentation: A valuable experience. *Medical Teacher* 20(6): 552-557.

Brooks-Brunn, Jo Ann. 1996. Poster etiquette. *Applied Nursing Research* 9(2): 97-99.

Carter-Thomas, Shirley, and Elizabeth Rowley-Jolivet. 2008. If-conditionals in medical discourse: From theory to disciplinary practice. *Journal of English for Academic Purposes* 7: 191-205.

Demarteau, Nadia, Karen Moeremans, and Lieven Annemans. 2007. Critical appraisal of scientific posters comparing anemia treatments for cancer patients: Applying ISPOR task force guidelines on methodological quality of retrospective studies. *Critical Reviews in Oncology/Hematology* 63: 91-99.

De Castro, Paola. 2009. *Librarians of Babel. A toolkit for effective communication.* Oxford: Chandos.

De Simone, Raffaele, Jörg Rodrian, Brigitte Osswald, Falk-Udo Sack, Eliana De Simone, and Siegfried Hagl. 2001. Letter to the editor. Initial experience with a new communication tool: The 'digital interactive poster presentation'. *European Journal of Cardio-Thoracic Surgery* 19: 953-955.

Di Blasio, Norina W. 2004. La comunicazione congressuale. Il poster. In *Diciamolo chiaramente*, ed. Paola De Castro, Silvana Guida and Bianca Maria Sagone, 229-239. Rome: Il Pensiero Scientifico Editore.

Dogan Bozdag, Ali. 2008. A new technique for presentation of scientific works: Video in poster. *World Journal of Surgery* 32: 1559-1561.

Dubois, Betty L. 1985a. Poster sessions at biomedical meetings: Design and presentation. *The ESP Journal* 4: 37-48.

—. 1985b. Popularization at the highest level: Poster sessions at biomedical meetings. *International Journal of the Sociology of Language* 56: 67-84.

Ellerbee, Susan. 2009. An artistic view of posters. *Trends in Neuroscience* 9(2): 109-110.

Erren, Thomas C., and Philip E. Bourne. 2007. Ten simple rules for a good poster presentation. *PLoS Computational Biology* 3(5): 777-778.

Facchinetti, Roberta. 2001. Conditionals constructions in Modern English legal texts. In *Modality in specialised texts*, ed. Maurizio Gotti and Marina Dossena, 133-150. Bern: Peter Lang.

Ferguson, Gibson. 2001. If you pop over there: A corpus-based study of conditionals in medical discourse. *English for Specifc Purposes* 20: 61-82.

Hay, Iain, and Susan M. Thomas. 1999. Making sense with posters in biological science education. *Journal of Biological Education* 33(4): 209-214.

Hess, George R., Kathryn W. Tosney, and Leon H. Liegel. 2009. Creating effective poster presentations: AMEE Guide no. 40. *Medical Teacher* 31(4): 356-358.

Huang, Stephen T., Maged N. Kamel Boulos, and Robert P. Dellavalle. 2008. Scientific discourse 2.0: Will your next poster session be in Second Life ®? *EMBO Reports* 9(6): 496-499.

Hunston, Susan. 2002. *Corpora in applied linguistics*. Cambridge: Cambridge University Press.

Keegan, David A., and Susan L. Bannister. 2003. Effect of colour coordination of attire with poster presentation on poster popularity. *Canadian Medical Association* 169(12): 1291-1292.

Lorenzoni, Paulo José, Raquel Canzi Almada de Souza, Suely Keiko Kohara, João César Beenke França, Giovanna Assis Rodrigues, and José Gastão Rocha de Carvalho. 2007. O pôster em encontros científicos. *Revista Brasileira de Educação Médica* 31(3): 304-309.

MacIntosh-Murray, Anu. 2007. Poster presentations as a genre in knowledge communication: A case study of forms, norms, and values. *Science Communication* 28(3): 347-376.

Matthews, Janice R., and Robert W. Matthews. 2007. *Successful scientific writing*. Cambridge: Cambridge University Press.

Maugh, Thomas H. 1974. Speaking of science: Poster sessions: A new look at scientific meetings. *Science* 184(4144): 1361.

Miller, Jane E. 2007. Preparing and presenting effective research posters. *HSR: Health Services Research* 42(1): 311-328.

Miracle, Vickie A. 2003. How to do an effective poster presentation in the workplace. *Dimensions of Critical Care Nursing* 22(4): 171-172.

Pearce, Euan L.F. 1992 Proceedings, IADR Council Meeting. Committee reports. *Journal of Dental Research* 71(10): 1677-1681.

Powell-Tuck, Jau, S. Leach, and L. MacCready. 2002. Electronic poster presentations in BAPEN: A controlled evaluation. *Clinical Nutrition* 21(3): 261-63.

Purrington, Colin. 2009. *Designing conference posters*. Available at http://colinpurrington.com/tips/academic/posterdesign. [May 10, 2014]

Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik. 1985. *A comprehensive grammar of the English language.* London: Longman.

Rowe, Nicholas, and Dragan Ilic. 2009. What impact do posters have on academic knowledge transfer? A pilot survey on author attitudes and experiences. *BMC Medical Education* 9(71). Available at www.biomedcentral.com/1472-6920/9/71 [May 10, 2014].

Rowley-Jolivet, Elizabeth. 2002. Visual discourse in scientific conference papers: A genre-based study. *English for Specific Purposes* 21(1): 19-40.

Saeed, Aziz T. 2004. Some pragmatic considerations in the positioning of *if*-clause in conditional sentences. *Studia Anglica Posnaniensia* 40: 245-255.

Smith, Philip E. M., Geraint Fuller, and Frank Dunstan. 2004. Scoring posters at scientific meetings: First impressions count. *Journal of the Royal Society of Medicine* 97: 340-341.

Stoss, Fred. 2010. *Designing effective poster presentations*. Available at http://ublib.buffalo.edu/libraries/asl/guides/bio/posters.html [May 10, 2014].

Swales, John M. 2004. *Research genres. Explorations and applications*. Cambridge: Cambridge University Press.

Swales, John M., and Christine B. Feak. 2000. *English in today's research settings*. Cambridge: Cambridge University Press.

van Naerssen, Margaret. 1984. Science conference poster presentations in an ESP program. *The ESPJournal* 3: 47-52.

Willet, Lisa L., Anuradha Paranjape, and Carlos Estrada. 2008. Identifying key components for an effective case report poster: An observational study. *Journal of General Internal Medicine* 24(3): 393-397.

Woolsey, John D. 1989. Combating poster fatigue: How to use visual grammar and analysis to effect better visual communication. *Transactions in Neurosciences (TINS)* 12(9): 325-332.

Zandifar, Ali, Ramani Duraiswami, and Larry S. Davis. 2005. A video-based framework for the analysis of presentations/posters. *IJDAR. International Journal on Document Analysis and Recognition* 7: 178-187.

# CHAPTER SIX

# TEXT REFLEXIVITY IN ACADEMIC WRITING: A CROSS-DISCIPLINARY AND CROSS-GENERIC ANALYSIS[1]

GIULIANA DIANI
UNIVERSITÀ DI MODENA E REGGIO EMILIA, ITALY

## 1. Introduction

Text reflexivity or "metatext" (Mauranen 1993) has been shown to play a significant role in academic discourse (e.g. Crismore 1983; Crismore and Farnsworth 1990; Mauranen 1993; Hyland 1998, 2005a; Dahl 2004a; Ädel 2006). Metatextual elements, that show how the text is organised or functionally structured or that make reference to the text itself and its units, can be seen to highlight the specificity of different genres, especially when signalling features of typical textual patterns or generic structures. Studies of academic discourse have shown that research genres are characterized by elements that point at the inherently metadiscursive nature of research discourse (e.g. Bondi 2001; Dahl 2004b; Hyland 2005b; Diani 2007; Ädel and Mauranen 2010).

Within EAP, most research has been devoted to the study of language variation across disciplines (e.g. Hyland 2000; Hyland and Bondi 2006; Fløttum 2007; Woodward-Kron 2008). Disciplinary discourses are often characterized by metadiscourse strategies as well as by content, and the so-called "general academic language" (Hyland and Bondi 2006) may in fact vary in discourse, in such a way as to reflect the epistemology of the discipline or the communicative peculiarities of the genre (Bondi 2014).

This chapter presents results of a study on the use of reflexive phraseology that characterizes two written academic genres in English: the research article and the book review article. Employing a corpus-based approach, the study focuses on how phraseological units vary across academic disciplines (with their variety of languages) and genres (with

their variety of communicative purposes and functions), thus following a recent interest not only in phraseology itself, but also in the phraseological features of academic discourse (e.g. Biber 2004; Biber et al. 2004; Groom 2005; Siepmann 2005; Charles 2006; Nesi and Basturkmen 2006; Hyland 2008a; Bondi 2010a).

The study presented in this chapter focuses on closely-related fields like business and economics. The choice of the two very 'close' disciplinary areas is in line with an interest in these closer distinctions, as well as with the need to better understand both fields. A considerable amount of research in the field of applied linguistics has focused on economics discourse in general (e.g. Merlini Barbaresi 1983; Tadros 1985; Henderson and Hewings 1987; Dudley-Evans and Henderson 1990; Henderson et al. 1993; Bondi 1999; Dahl 2009), but little attention has been paid to the distinction between business and economics discourse (Hemais 2001; Bondi 2006, 2010b). Focusing on the role of narrative elements in business and economics, Bondi (2006) shows that forms of narrative from the world of business are often used as illustrations, examples or case studies that provide the main argumentative line of the article, as marked by key signals of narrative development. The "rhetoric of economics", on the other hand, can be shown to rely on abstract reasoning and argument (Bondi 1999). As Bondi explains (1999, 71-72):

> Economic reasoning is often presented in terms of possible worlds: discourse follows a logic of ramification determining the range of possibilities that emerge from a state of affairs. Economics as a science proceeds not only by representing what happens in the actual world, but also by exploring what could have or could not have occurred in actuality. Scientific argument in economics can hardly be based on laboratory experiment: its domain is a comparative examination of situations taken from the world of fact or in forms of reasoning based on models of reality.

The choice of the written academic genres under examination – the research article and the book review article – is linked to their very specific status in the field of genre studies. Keeping in mind the basically dialogic and argumentative nature of academic discourse, both the book review article and the research article represent the most distinguished channel of knowledge dissemination within the specific scientific community. Within an academic context, they play a crucial role in the process of knowledge construction and discussion by providing a forum in which academics can set out their views in the form of arguments. They

are thus most likely to provide signals of their preferred reasoning structures.

In this study I attempt to shed light on the way academics in the field of business and economics represent their own argument in written academic genres like research articles and book review articles through the analysis of reflexive phraseological elements. More specifically, I aim to answer the following questions: does the reflexive phraseology reflect disciplinary variation across closely-related fields like business and economics? Does it reflect the different communicative purposes of equally argumentative genres like the research article and the book review article?

After detailing materials and methods in Section 2, I focus on reflexive phraseology ranging from comparison of frequency data to contextual analysis of phraseological units across disciplines (Section 3, on business and economics). The elements thus highlighted are further explored across academic genres (Section 4, on economics research articles and book review articles).

## 2. Materials and methods

This study is based on the analysis of three specialised corpora of research articles and book review articles, which have been designed to study academic writing in English across genres and disciplines. I made use of the following corpora:

a) a corpus of 436 economics research articles (HEM-Economics) published in eight British and American academic journals spanning the years 1999-2000 (consisting of about 2,500,000 words).[2]
b) a corpus of 370 business research articles (HEM-Marketing) published in seven British and American business academic journals spanning the years 1999-2000 (consisting of about 2,500,000 words).[3]
c) a corpus of 24 economics book review articles (EBRA) published in five British and American economics academic journals spanning the years 2000-2003 (consisting of about 167,000 words).[4]

The corpus of economics book review articles is of a different size because this reflects a much more limited presence in academic genres. All

frequency data reported in this study will be presented as normalized figures, calculated per ten thousand words.

The methodology adopted for this study combines a discourse and a corpus perspective. Discourse analysis contributes to the definition of pragmatic functions of the reflexive lexical items under investigation, whereas corpus linguistics offers ways of looking at lexical patterns: in particular, using the wordlist and keyword function of WordSmith Tools 5.0 (Scott 2007), I studied concordances, wordlists and lists of clusters (i.e. repeated strings of words, irrespective of structural units; see also Biber et al. 2004), as well as keywords and key-clusters. These were worked out by comparing corpora to each other.

The analysis carried out focuses on:

a) cross-disciplinary variation in the presence of lexical tools of reflexivity in economics and business research articles;
b) cross-generic variation of tools of reflexive language in research articles and book review articles in the field of economics discourse.

While starting from statistically-determined key-clusters, the study looks at phraseology on the basis of a combination of frequency-based information and semantics. The frequencies of word forms and "multiword units" (MWUs) (Moon 1997) provide a useful starting point for closer examination of extended lexical units (Sinclair 1996), with their corollary of semantic preference and typical pragmatic association. Key-clusters are also very helpful when focusing on disciplinary differences (Hyland 2008b; Malavasi and Mazzi 2010).

The present analysis focuses on "organizational discourse units" – "elements which organize upfolding discourse" (Sinclair and Mauranen 2006, 70) – playing a major role in the construction of economists' discourse: from general discourse markers (*on the other hand*, *as a result*) to other meta-argumentative and self-reflexive items (*it is shown that*, *the purpose of this paper*, *the results of this study*, *is organized as follows*).

## 3. Lexical tools of reflexivity across disciplines: Exploring frequency data

Key-clusters provide a useful starting point for the identification of typical phraseology in the corpora under investigation. A look at 4- or 5-word clusters highlights some of the most frequent phraseological units in the disciplinary corpora under investigation.

Table 1 below reports the ten most frequent 5-word clusters identified in the corpora. As the data show, they are mostly organisational. The expressions refer to the text itself (e.g. *the paper is organized as*), or to quantitative data presented in tables (e.g. *as shown in table #*, *are presented in table #*), with their mathematical or statistical aspects (e.g. *less than or equal to, significant at the # level*).

| | Business | | Economics | |
|---|---|---|---|---|
| 1 | on the basis of the | 81 | table # and table # | 184 |
| 2 | arithmetic mean of sample | 76 | paper is organized as follows | 102 |
| 3 | less than or equal to | 76 | significant at the # level | 97 |
| 4 | mean of sample # | 73 | at the end of the | 85 |
| 5 | are more likely to be | 65 | are presented in table # | 78 |
| 6 | at the end of the | 63 | as a function of the | 77 |
| 7 | the extent to which the | 62 | the paper is organized as | 77 |
| 8 | as shown in table # | 61 | at the beginning of the | 62 |
| 9 | in the context of the | 48 | of this paper is to | 60 |
| 10 | are shown in table # | 47 | as a result of the | 59 |

**Table 1. Ten most frequent 5-word clusters in business and economics.**

Tables 2 and 3 report examples and quantitative data of economics and business phraseology classified according to typical discourse function.

| | Functions | Expressions | Economics Freq. per 10,000 words | Business Freq. per 10,000 words |
|---|---|---|---|---|
| Meta-argumentative expressions | Introducing assumptions | *we assume that the* | 0.5 (133 occur.) | 0.1 (43 occur.) |
| | | *is assumed to be* | 0.5 (130) | 0.1 (22) |
| | | *it is assumed that* | 0.4 (114) | 0.1 (16) |
| | Introducing claims/ conclusions | *it is shown that* | 0.2 (49) | 0 |
| | | *it can be shown that* | 0.2 (48) | 6 occur. |
| | | *it is straightforward to show* | 0.1 (21) | 0 |
| | Introducing research structure | *the results of this paper* | 0.1 (20) | 0 |
| Metatextual expressions | Introducing purpose | *in this paper we* | 0.3 (70) | 12 occur. |
| | | *the purpose of this paper* | 0.2 (49) | 13 occur. |
| | Introducing textual structure | *is organized as follows* | 0.4 (113) | 0.1 (29) |
| | | *the remainder of the paper* | 0.1 (36) | 5 occur. |
| | | *the paper is as follows* | 0.1 (36) | 5 occur |
| | | *the second half of the* | 0.1 (21) | 3 occur. |
| | | *section # concludes the paper* | 0.1 (20) | 0 |
| | | *are presented in section #* | 0.1 (19) | 0 |

**Table 2. Key reflexive phraseology in economics (normalised per ten thousand words).**

| | Functions | Expressions | Business Freq. per 10,000 words | Economics Freq. per 10,000 words |
|---|---|---|---|---|
| Meta-argumentative expressions | Introducing assumptions | *on the basis of* | 1 (358 occur.) | 0.6 (156 occur.) |
| | Introducing research structure (analysing data) | *research has shown that* | 0.2 (47 occur.) | 4 |
| | | *will be positively related to/associated to* | 0.2 (45) | 0 |
| | | *the results of this study* | 0.2 (44) | 8 |
| | | *there were no significant* | 0.1 (35) | 0 |
| | | *support for hypothesis #* | 0.1 (31) | 0 |
| | | *there was no significant* | 0.1 (30) | 0 |
| Metatextual expressions | | *in the present study* | 0.2 (42) | 4 |

**Table 3. Key reflexive phraseology in business (normalised per ten thousand words).**

The analysis highlights an interesting element of disciplinary variation. Economics exhibits a greater preference for metatextual expressions (e.g. *the purpose of this paper*, *the paper is as follows*, *in this paper we*), whereas business shows a more frequent use of expressions signalling the process of research (e.g. *support for hypothesis*, *will be positively related to*) and the analysis of data (e.g. *the results of this study*). Obvious differences can also be noticed within the same function: in 'Introducing assumptions', for example, economics shows a preference for expressions centred around the verb *assume* followed by propositions, whereas business clearly favours prepositional constructs, where *on the basis of* is variously followed by *data*, *tools*, *theories*, *construct*.

As rightly noted by Bondi (2010b, 227), these formal differences can be explained in terms of methodological issues: "the key role played by mathematical reasoning and philosophical argument in economics, as against the central role of empirical research and theory testing in business studies".

# 4. Economics across genres

## 4.1. Focus on discourse markers

When considering general discourse markers across the two genres under examination, I find that the most frequent items fall into the semantic categories of 'contrast/concession' like *on the other hand*, *at the same time* and 'result/inference' like *as a result*. Although the figures are higher in research articles than in book review articles (see Table 4), a brief consideration of these pragmatic markers brings out the importance of the representation of the process of arguing in both genres analysed.

| Pragmatic markers in economics | Research articles (HEM-Economics) Freq. per 10,000 words | Book review articles (EBRA) Freq. per 10,000 words |
|---|---|---|
| Contrastive/concessive | | |
| - *on the one hand* | 0.4 (106 occurrences) | 0.3 (5 occurrences) |
| - *on the other hand* | 2.2 (561 occurrences) | 1.0 (18 occurrences) |
| - *at the same time* | 0.1 (22 occurrences) | 1.3 (23 occurrences) |
| Conditional-restrictive | | |
| - *in the case of* | 1.5 (388 occurrences) | 0.6 (11 occurrences) |
| - *with respect to the* | 0.6 (164 occurrences) | 0.6 (10 occurrences) |
| - *if and only if* | 0.6 (150 occurrences) | 1 occurrence |
| Causal | | |
| - *as a result* | 0.7 (185 occurrences) | 0.5 (9 occurrences) |

**Table 4. Pragmatic markers in economics research articles and book review articles.**

Both book review articles and research articles are shown to be particularly interested in representing the argumentative procedures of the community: the various lexicalisations are used in the sense of introducing positions with their objections or concessions.

Examples (1) and (2) taken from EBRA and HEM-Economics illustrate moves in which we are offered a representation of debate within the discipline. Notice the expressions referring to argument evaluation, where divergence of opinions is emphasised by the use of *on the other hand*: (1 – *On the other hand, we explained before that option model based forecast requires a number of assumptions […] to produce a useful volatility estimate […]*; 2 – *On the other hand, the question how monitoring takes place or should take place is not answered satisfactorily,*

*yet*). The discourse marker signals a divergence in the World of Discourse (in the interpretation of economic processes) rather than the World of Text (in the economic processes themselves), in line with a general proclivity of economics discourse for framing argument within the context of the scientific community.

**(1)**
As mentioned earlier, option implied volatility is perceived as a market's expectation of future volatility and hence it is a market based volatility forecast. Arguably it should be superior to a time series volatility forecast. **On the other hand**, we explained before that option model based forecast requires a number of assumptions to hold for the option theory to produce a useful volatility estimate. […] (*EBRA*)

**(2)**
Why is there so much monitoring? Whenever a principal delegates a task to an agent who is better informed and has different interests than the principal, the agent may take advantage of the asymmetry of information. […] The principal can do the monitoring herself or she can send an auditor. So it is well understood why there is so much auditing. **On the other hand**, the question how monitoring takes place or should take place is not answered satisfactorily, yet. How often should monitoring take place? How thorough should the individual audit be? How should a convicted agent be punished? What does optimal monitoring look like, if auditor and agent may collude? In this paper, I will provide some new ideas and insights regarding these questions. (*HEM-Economics*)

The contrastive pattern exemplified above not only represents the most important sequence in the construction of argument in both genres analysed, but also the most important pattern in which the adverbial *on the other hand* is involved.

When focusing on the use of the contrastive markers in the corpora, we may notice a considerable difference between cases. If the adverbial *on the other hand* in examples (1) and (2) above is mostly used for contrasting two arguments which may lead to different conclusions (1) or which require further argument to be turned into a coherent whole (2), the adverbial *at the same time* in example (3) is used to contrast two propositions which are presented as compatible, but opposing in pragmatic function (praise and criticism):

**(3)**
The book presents sensible and intuitive behavioral stories dealing with an impressive range of financial phenomena, and then supports these

> stories with a vast amount of compelling empirical and anecdotal evidence. **At the same time**, however, the book only goes part of the way to making the case that the general behavioral finance program presents a better way to do research in finance. Missing from such an argument is any notion that the field as a whole can provide clear empirical hypotheses, or that the field is moving towards any coherent framework that can provide organization and structure to its many disparate ideas. (*EBRA*)

The balancing of alternative scenarios is even more explicit when *on the other hand* is accompanied by a formulaic use of its correlative adverbial *on the one hand*, as illustrated in example (4).

> **(4)**
> Grossman and Helpman do not study coalitions among SIGs, but their models provide a starting point for such an analysis. **On the one hand**, when cheap-talk lobbying is the instrument of influence, SIGs located on the same side of the policymaker are unable to communicate more information than what the most moderate SIG can communicate. In such a case, the incentives to form a coalition may be weak. **On the other hand**, when contributions are the instrument of influence, SIGs with aligned policy interests may have an incentive to form a coalition. (*EBRA*)

The extract above shows that the reviewer does not react to the theory under review in a completely negative or completely positive way. Rather, he weighs up the pros and cons of it. This seems to indicate that his opinion is not one-sided and that he has given some consideration to both sides of a question. This strategy assumes a turning role: it may serve to create a more balanced comment, slightly softening the negativity of the evaluation. In research articles, the occurrence of similar patterns, with both markers, constructs the argument as balanced right from the beginning, highlighting that different options and features have been considered.

Among the various sequences that involve argument and debate on the part of the researcher and the reviewer, the most frequent in both corpora contain items that signal conditional/restrictive and casual relationships like *as a result*. The example below illustrates this point.

> **(5)**
> […] sophisticated attempts to establish the direction of causality of a growth relationship, either through instrumentation or the use of time lags, are rarely based on structured models of the process being estimated (rather, they are based on finding "clever" instruments). **As a**

> **result**, the corresponding estimates are hard to interpret, and these attempts are generally characterized by heroic claims of causality—this is a topic to which we shall return below. So what are the main correlates of economic growth? Research has focused on several broad categories of growth determinants, and rather than exhaustively reviewing these numerous determinants, we will focus on the categories. (*EBRA*)

The first sentence of the extract builds up a claim about an economic topic (*sophisticated attempts to establish the direction of causality of a growth relationship […] are rarely based on structured models […]*). The adverbial *as a result* introduces the consequence of the preceding discourse, signalling both the reviewer's opinion and the conclusions that the reviewer expects the reader to draw (*As a result, the corresponding estimates are hard to interpret, and these attempts are generally characterized by heroic claims of causality*).

Although both genres use conditional/restrictive markers like *in the case of*, *with respect to the*, *if and only if* for similar purposes, i.e. allowing the researcher or the reviewer to specify the conditions under which argument holds, book review articles place less emphasis on conveying logical coherence and building arguments, which is reflected in a lower frequency: 11 (0.6 pttw) occurrences of *in the case of* in EBRA when compared with 388 (1.5 pttw) occurrences in HEM-Economics, and only 1 (0.05 pttw) occurrence of *if and only if* in EBRA against 150 (0.6 pttw) occurrences in HEM-Economics.

## 4.2. Focus on meta-argumentative expressions across genres

We move on now to the analysis of meta-argumentative expressions across the corpora. My main interest is in the lexicalisations of argumentative procedures, when they are used to introduce assumptions.

Economics has been shown to make extensive use of signals of assumption and hypotheticality (Tadros 1985; Bondi 1999). The most interesting cases of references to the world of hypothesis are those where a hypothetical case is made to provide support for a general claim made about the real world. The most prominent pragmatic function of hypothetical moves thus consists in providing examples or illustrations for the generalisations put forward by the writer.

When considering the lemma *assume*, economics research articles manifest a clear preference for relating use of it with the first person plural pronoun like *we assume that the* (133 occurrences, 0.5 pttw), followed by the patterns *is assumed to be* (130 occurrences, 0.5 pttw) and *it is assumed*

*that* (114 occurrences, 0.4 pttw). Economics book review articles, on the other hand, show slightly different patterns. They are more limited in the range of references to the world of hypothesis. They are shown to relate use of *assume* with first person forms in only two cases out of 35 occurrences in the corpus: *I will assume that*…; *we assume that*… .The pattern *it is assumed that* occurs only 3 times. The overwhelming majority of its occurrences are used as reporting expressions with human subjects (*they assume that*…). This result can be explained by the tendency for book review articles to focus on a representation of the discursive procedures of the discourse community: in particular, a representation of what the reviewed author says. It is characteristic of the genre under examination that the reviewer manifests his/her identity presenting and evaluating the reviewed author's work through constant references to the reviewed author's ideas. Reporting the reviewed author's claims contributes to representing the debate the reviewer builds with the reviewed author and the various participants in the dialogue.

Further differences can be noticed in comparing frequencies across genres. A search for reporting expressions in the clause type 'introductory *it*' followed by passive voice used for introducing the reviewer's or researcher's claims and conclusions like *it is shown that* or *it can be shown that* revealed very few occurrences in the EBRA corpus: only 2 occurrences of *it can be shown that* were found, whereas they are frequent in economics research articles (*it is shown that* 49 occurrences; *it can be shown that* 48; *it is straightforward to show* 21). A search for other reporting verbs in that clause type in the EBRA corpus still produced only 8 relevant occurrences: *it is generally recognized that* (3 occurrences)*, it is well known that* (3)*, it is said that* (1)*, it is argued that* (1).

If we compare the frequencies of the verb *show* in these clause types with those clauses with human subjects (first person singular/plural pronouns or third persons), evidence of further patterning emerges, as shown in Table 5.

| *Show* in two clause types | Economics research articles (HEM-Economics) Freq. per 10,000 words | Economics book review articles (EBRA) Freq. per 10,000 words |
|---|---|---|
| 'introductory it'+ passive voice | | |
| *It is shown that* | 0.2 (49 occurrences) | 0 |
| *It can be shown that* | 0.2 (48 occurrences) | 2 occurrences |
| *It is straightforward to show* | 0.1 (21 occurrences) | 0 |
| 1st person singular/plural pronouns | | |
| *I + show* | 0.1 (29 occurrences) | 2 occurrences |
| *we + show* | 0.5 (118 occurrences) | 0 |
| 3rd person singular/plural pronoun and specific author names | 0.8 (202 occurrences) | 7.9 (132 occurrences) |

**Table 5. Frequencies of the lemma *show* in clauses used for introducing claims and conclusions.**

In economics research articles researchers tend to introduce their claims both through the use of 1st person pronouns and passive voice. This may find an explanation in their role in writing the article, i.e. communicating results. The very clear signalling of the researcher's activity is given by the high frequency of the verb *show* in both clause types ('it' subject with passive and human subject). The frequency of occurrence of the verb is similar in both clauses: 0.5 pttw in 'introductory it' + passive voice; 0.6 pttw in human subject clause type. The frequent use of *show* in the genre "reflects both the importance of the researcher's investigation and the concern with discovery" (Bondi 2010b, 226). In book review articles, on the other hand, the verb *show* mostly occurs in clauses with references to the reviewed author.

We may explain these differences by reference to differences in the principal rhetorical purposes of the two genres. Research articles primarily function to construct and promote new knowledge claims (Bazerman 1988; Swales 1990), while book review articles primarily function to evaluate and discuss the knowledge claims of other researchers (Hyland and Diani 2009).

## 4.3. Focus on meta-textual expressions across genres

As has been shown in previous studies (Diani 2004, 2007), the structure of a book review article is fairly similar to that of a research article. It is divided into the traditional sections: Introduction, Critique, Conclusion. The Introduction and the Conclusion are often very similar in

structure to those of a research article, although they have to include a specific focus on the presentation and final evaluation of the work they review. The Critique section, representing the body of the article, is usually characterized by a highly cyclical structure: each single issue identified is discussed critically in a series of sections.

The aim here is to focus on cross-generic comparison of lexical tools introducing generic rhetorical structure in the typical move of 'Introduction'.

If we consider the lexical items used by both reviewers and researchers to present their purpose in writing the article, undoubtedly the most frequent (albeit predictable) expressions are related to deictic elements (*this*), and self-referential nouns referring to the discourse product itself (*paper, review, section*). Characteristic patterns can be seen below, in example (6) from the EBRA corpus.

**(6)**
**In this paper, I briefly review s**ome of the main issues which this collection deals with, and suggest directions for future research. **Section 2 looks** at the age-old question of why Britain repealed the Corn Laws in 1846, and maintained a free-trading stance for the rest of the century, while **Section 3 deals with** the consequences of Britain's free trade commitment. **Section 4 concludes**. (*EBRA*)

In economics research articles the most frequent pattern for introducing the purpose and structure of the article is by far *the paper is organized as follows* (113 occurrences), followed by *the remainder of the paper* (36), *the second half of the* (31), *the paper is as follows* (21). The search for nominal references to the notion of "purpose" (Bondi 2007) produced 49 occurrences in the pattern *the purpose of this paper*.

Moving on to economics book review articles, I find that the corpus produced a more varied picture. The organisational structure of the article is mostly expressed through the combination of first person singular and plural pronouns (26 relevant hits) and a range of verbs like *review*, *focus*, *discuss*, *examine, describe, illustrate, summarise, conclude*, as in (6) above. When we analyse nominal references to the notion of purpose, I find a picture of greater divergence compared with HEM-Economics: no occurrence of nouns referring to the notion was found in the corpus.

On the whole, nominal references to the discourse unit prove to be the most frequent choice in HEM-Economics, whereas they are far more limited in EBRA: I found only one occurrence of the pattern *the remaining sections are organized as follows*.

## 5. Conclusions

This analysis of reflexive phraseology across disciplines and genres has revealed that organisational phraseological units can contribute to the study of textual patterns or generic structures. Whether they act at the level of discourse and interaction or at the level of the text and its linear structure, organisational phraseological units are shown to play a central role in academic discourse: "outlining assumptions, problematizing issues, highlighting the significance of the data and the conclusions produced" (Bondi 2010b, 231).

Organisational units also point at key nodes in the structure of genres, highlighting for example the different elements of scenario-based formal reasoning in research papers and the interplay between praise and criticism in book reviewing.

Similarities and differences between closely related disciplines and genres may also help focus on disciplinary discourse. The meta-argumentative phraseological elements that characterize both research articles and book review articles can be considered, to borrow Bondi's (2010b, 231) words, as "a resource by which the author negotiates his/her position with the reader according to discipline-specific orientations, showing for example emphasis on abstract reasoning in economics or emphasis on causal sequences in accounting for factual data in business studies".

## Notes

[1] The source of this chapter is Bondi and Diani (2008).

[2] The journals considered are: *European Economic Review, European Journal of Political Economy, International Journal of Industrial Organization, International Review of Economics and Finance, Journal of Corporate Finance, Journal of Development Economics, Journal of Socio-Economics, The North American Journal of Economics and Finance.*

[3] The journals considered are: *Journal of World Business, Academy of Management Journal, Marketing Science, Journal of Marketing Research, Business Economics, Business & Society Review, Business Strategy Review*.

[4] The journals considered are: *Journal of Economic Literature, European Journal of Political Economy, Information Economics and Policy, Journal of Economics, Structural Change and Economic Dynamics, Journal of Monetary Economics*.

## References

Ädel, Annelie. 2006. *Metadiscourse in L1 and L2 English*. Amsterdam: John Benjamins.

Ädel, Annelie and Anna Mauranen. eds. 2010. Metadiscourse. Special issue of the *Nordic Journal of English Studies* 9(2).

Bazerman, Charles. 1988. *Shaping written knowledge*. Madison, Wisconsin: The University of Wisconsin Press.

Biber, Douglas. 2004. Lexical bundles in academic speech and writing. In *Practical applications in language and computers* (PALC 2003), ed. Barbara Lewandowska-Tomaszczyk, 165-178. Frankfurt am Main: Peter Lang.

Biber, Douglas, Susan Conrad, and Vivian Cortes. 2004. "It you look at…": Lexical bundles in university teaching and textbooks. *Applied Linguistics* 25(3): 371-405.

Bondi, Marina. 1999. *English across genres: Language variation in the discourse of economics*. Modena: Il Fiorino.

—. 2001. Small corpora and language variation: Reflexivity across genres. In *Small corpus studies and ELT. Theory and practice*, ed. Mohsen Ghadessy, Alex Henry and Robert L. Roseberry, 135-174. Amsterdam: John Benjamins.

—. 2006. "A case in point": Signals of narrative development in business and economics. In *Academic discourse across disciplines*, ed. Ken Hyland and Marina Bondi, 47-72. Bern: Peter Lang.

—. 2007. Historical research articles in English and in Italian: A cross-cultural analysis of self-reference in openings. In *Lexical complexity: Theoretical assessment and translational perspectives*, ed. Marcella Bertuccelli Papi, Gloria Cappelli and Silvia Masi, 65-84. Pisa: PLUS.

—. 2010a. Abstract writing: The phraseology of self-representation. In *Linguistic interaction in/and specific discourses*, ed. Marta Conejero Lopez, Micaela Muñoz Calvo and Beatriz Penas Ibáñez, 31-48. Valencia: Universitat Politècnica de Valencia.

—. 2010b. Arguing in economics and business discourse: Phraseological tools in research articles. *Bulletin Suisse de Linguistique Appliquée* 2: 219- 234.

—. 2014. Connecting science. Organizational units in specialist and non-specialist discourse. In *The language of popularization: Theoretical and descriptive models / Die sprache der popularisierung: Theoretische und deskriptive modelle*, ed. Giuditta Caliendo and Giancarmine Bongo, 51-72. Bern: Peter Lang.

Bondi, Marina, and Giuliana Diani. 2008. Forms of metadiscourse in English academic writing: A cross-disciplinary and cross-generic analysis of meta-argumentative phraseology. In *Threads in the complex fabric of language. Linguistic and literary studies in honour of Lavinia Merlini Barbaresi*, ed. Marcella Bertuccelli Papi, Antonio Bertacca and Silvia Bruti, 69-84. Pisa: Felici Editore.

Charles, Maggie. 2006. Phraseological patterns in reporting clauses used in citation: A corpus-based study of theses in two disciplines. *English for Specific Purposes* 25(3): 310-331.

Crismore, Avon. 1983. *Metadiscourse: What it is and how it is used in school and non-school social science texts*. Urbana-Champaign, Center for the Study of Reading: University of Illinois.

Crismore, Avon, and Rodney Farnsworth. 1990. Metadiscourse in popular and professional science discourse. In *The writing scholar*: *Studies in academic discourse*, ed. Walter Nash, 118-136. Newbury Park: Sage.

Dahl, Trine. 2004a. Textual metadiscourse in research articles: A marker of national or of academic discipline? *Journal of Pragmatics* 36(10): 1807-1825.

—. 2004b. Some characteristics of argumentative abstracts. *Akademisk Prosa* 2: 49-67.

—. 2009. The linguistic representation of rhetorical function: A study of how economists present their knowledge claims. *Written Communication* 26(4): 370-391.

Diani, Giuliana. 2004. A genre-based approach to analysing academic review articles. In *Academic discourse, genre and small corpora*, ed. Marina Bondi, Laura Gavioli and Marc Silver, 105-126. Rome: Officina Edizioni.

—. 2007. The representation of evaluative and argumentative procedures: Examples from the academic book review article. *Textus* 20(1): 37-56.

Dudley-Evans, Tony, and Willie Henderson. eds. 1990. *The language of economics. The analysis of economic discourse*. London: Modern English Publications/British Council.

Fløttum, Kjersti. ed. 2007. *Language and discipline perspectives on academic discourse*. Newcastle: Cambridge Scholars Publishing.

Groom, Nicholas. 2005. Pattern and meaning across genres and disciplines: An exploratory study. *Journal of English for Academic Purposes* 4(3): 257-277.

Hemais, Barbara. 2001. The discourse of research and practice in marketing journals. *English for Specific Purposes* 2: 39-59.

Henderson, Willie, Tony Dudley-Evans, and Roger Backhouse. eds. 1993. *Economics and language*. London: Routledge.

Henderson, Willie, and Ann Hewings. 1987. *Reading economics. How text helps or hinders*. British National Bibliography Research Fund, Report 28.

Hyland, Ken. 1998. Persuasion and context: The pragmatics of academic metadiscourse. *Journal of Pragmatics* 30(4): 437-455.

—. 2000. *Disciplinary discourses: Social interactions in academic writing*. London: Longman.

—. 2005a. *Metadiscourse*. *Exploring interaction in writing*. London: Continuum.

—. 2005b. A convincing argument: Corpus analysis and academic persuasion. In *Discourse in the professions: Perspectives from corpus linguistics*, ed. Ulla Connor and Thomas A. Upton, 87-114. Amsterdam: John Benjamins.

—. 2008a. Academic clusters*:* Text patterning in published and postgraduate writing. *International Journal of Applied Linguistics* 18(1): 41-62.

—. 2008b. "As can be seen": Lexical bundles and disciplinary variation. *English for Specific Purposes* 27: 4-21.

Hyland, Ken, and Marina Bondi. eds. 2006. *Academic discourse across disciplines*. Bern: Peter Lang.

Hyland, Ken, and Giuliana Diani. eds. 2009. *Academic evaluation: Review genres in university settings*. Basingstoke: Palgrave Macmillan.

Malavasi, Donatella, and Davide Mazzi. 2010. History v. marketing: Keywords as a clue to disciplinary epistemology. In *Keyness in texts*, ed. Marina Bondi and Mike Scott, 169-184. Amsterdam: John Benjamins.

Mauranen, Anna. 1993. Contrastive ESP rhetoric: Metatext in Finish-English economics texts. *English for Specific Purposes* 12(1): 3-22.

Merlini Barbaresi, Lavinia. 1983. *Gli atti del discorso economico: La previsione*. Parma: Edizioni Zara.

Moon, Rosamund. 1997. Vocabulary connections: Multi-word items in English. In *Vocabulary: Description, acquisition and pedagogy*, ed. Norbert Schmitt and Michael McCarthy, 40-63. Cambridge: Cambridge University Press.

Nesi, Hilary, and Helen Basturkmen. 2006. Lexical bundles and discourse signalling in academic lectures. *International Journal of Corpus Linguistics* 11(3): 283-304.

Scott, Mike. 2007. *WordSmith Tools*. Oxford: Oxford University Press.

Siepmann, Dirk. 2005. *Discourse markers across languages. A contrastive study of second-level discourse markers in native and non-native text*

*with implications for general pedagogy and lexicography*. London/New York: Routledge.
Sinclair, John McH. 1996. The search for units of meaning. *Textus* 9(1): 75-106.
Sinclair, John McH., and Anna Mauranen. 2006. *Linear unit grammar. Integrating speech and writing*. Amsterdam: John Benjamins.
Swales, John. 1990. *Genre analysis. English in academic and research settings*. Cambridge: Cambridge University Press.
Tadros, Angela. 1985. *Prediction in text*. Birmingham: English Language Research Monograph.
Woodward-Kron, Robyn. 2008. More than just jargon: The nature and role of specialist language in learning disciplinary knowledge. *Journal of English for Academic Purposes* 7(4): 234-249.

# PART II

# CONTRASTIVE EAP RHETORIC

# CHAPTER SEVEN

# INTERCULTURALITY IN EAP RESEARCH: PROPOSALS, EXPERIENCES, APPLICATIONS AND LIMITATIONS[1]

ROSA LORÉS SANZ
UNIVERSIDAD DE ZARAGOZA, SPAIN

## 1. Why do cross-cultural research?

It is by now a proven fact that academics face an increasing pressure to disseminate their research internationally in English. This has reinforced the hegemonic role of English as the language in which scientific knowledge is created and disseminated. The fact that one single language has become the vehicle of scientific communication has obvious advantages, as it facilitates and assures knowledge dissemination. However, it may be felt as an extra burden for non-anglophone academics, aware as they are that the recognition of their research by their disciplinary community very much depends on how proficient and successful they are in English for international communication.

Together with this, it is also generally accepted that our native linguistic culture has an impact on the drafting of academic texts, not only at a lexico-grammatical level, but, most importantly, at a rhetorical and discursive level. Interestingly, the challenges and difficulties that the use of EAP (English for Academic Purposes) as an L2 entails for non-Anglophone speakers have involved a reinforcement of the role of Intercultural Rhetoric (IR) (Connor 2004a, 2004b) in the study of English as the language of international academic and scientific communication, as well as a diversification of its methods and applications.

This chapter attempts to present and illustrate the methodological approach adopted by the research group InterLAE (www.interlae.com) for the intercultural study of academic genres written in English by non-native (Spanish) academics, which involves corpus analysis, genre analysis and

intercultural rhetoric, and the design, construction and analysis of the SERAC corpus (Spanish-English Research Article Corpus), as a tool for the analysis of interculturality within EAP. Moreover, some of the findings resulting from the study of SERAC about non-native (Spanish) use of EAP will be presented. Special emphasis will also be made on the assets and the limitations faced by this methodological approach. Finally, the advantageous adoption of a cross-cultural approach to the design of teaching materials and the implementation of EAP courses will also be discussed.

## 2. Background. From Kaplan to Connor

The term *Contrastive Rhetoric* goes back to Robert Kaplan's 1966 article "Cultural thought patterns in intercultural education". Kaplan explored the "pedagogical needs" of American teachers teaching composition to foreign students. As he later claimed (1988, 277), these teachers "were able to tell with astonishing accuracy what the native-language of the writer was". This made Kaplan suggest that there were some regularities in the ways foreign students from certain linguistic backgrounds wrote in English, as well as in the way native English speakers write. Kaplan's aim was to help the foreign students in their writing tasks, so it seemed logical to find out where their writing deviated from that of English native speakers. The deviations Kaplan was thinking of were at the rhetorical level, rhetoric understood as the choice of linguistic and structural aspects of discourse chosen to produce an effect on an audience (Purves 1988, 9). Thus, the ultimate objective of research activity in Contrastive Rhetoric has been to prove the hypothesis that different languages have different rhetorical systems which manifest themselves in different ways of organizing and developing ideas. The second half of the 20$^{th}$ century was its heyday and it has brought into existence a mass of contrastive studies conducted in many parts of the world, involving many languages, implementing many linguistic frameworks, and covering many aspects of language.

Kaplan's proposal was clearly influenced by the Sapir-Whorf hypothesis, according to which languages are shaped by our view of the world, that is, by our culture. Thus, the concept of "culture" plays a primary role in the understanding of the workings of Contrastive Rhetoric.

It is indisputable that culture influences writing habits in an important way. According to Mauranen (1993, 4) this is "because writing clearly is a cultural object, existing only in the social world of humans, as a product of social activities". However, as she points out, certain areas of culture are

generally assumed to be universal in a way that renders cultural variation unimportant. She specifically mentions the case of science: the phenomenon of scientific thought is seen as emerging from a universal source. But science, or more specifically, academic research, does not exist outside writing. Science is "realised" in language, and mainly in written form. Therefore, "we cannot represent it, or realise it, without being influenced by the variation in the writing cultures that carry it" (Mauranen 1993, 4).

Atkinson (2004) considers that the complexity of the culture stems from its being a dynamic construct (culture as process) which forces us to see it in a constant evolving perspective. Based on Holliday (1999), Atkinson establishes an interesting distinction between what he calls *big cultures* (national cultures) and *small cultures* (e.g. professional-academic cultures). It is interesting to notice that there is a certain degree of overlap between big and small cultures. The conceptualization of culture in terms of big and small, with the corresponding areas of overlap, provide a much more complex but also realistic picture of the interaction of different cultural forces and how this interaction is manifested in language.

In the 2004 monographic issue on Contrastive Rhetoric of the *Journal of English for Academic Purposes*, Ulla Connor (2004a) points out the negative connotation that the term *contrastive rhetoric* has gained over the years. As she states (2004a, 272), "contrastive rhetoric is often characterized as static, and is linked to contrastive analysis, a movement associated with structural linguistics and behaviourism". Thus, to distinguish between this "static" model and the new advances that have been made, she proposes the use of the term *intercultural rhetoric* to refer to the "current dynamic models of cross-cultural research" (2004a, 272). In her view, the term includes the majority of text and genre analysis studies that are now being carried out and also allows for the study of interactions in intercultural settings consisting of oral and written texts. Connor (2004a, 273) justifies the new proposal in the following terms:

> [Intercultural rhetoric] better describes the broadening trends of writing across languages and cultures. It preserves the traditional approaches that use text analysis, genre analysis, and corpus analysis as well as introduces the ethnographic approaches that examine language in interactions. Furthermore, it connotes the analysis of texts that allows for dynamic definitions of culture and the inclusion of smaller cultures (e.g. disciplinary, classroom) in the analysis.

As Connor, Nagelhout and Rozycki (2008, 3) state in the Introduction to their edited book *Contrastive Rhetoric: Reaching to Intercultural*

*Rhetoric*, many new trends have appeared in research and methods which explain the significant changes that have taken place in contrastive rhetoric. These changes respond to two major developments: the fact that more genres with specific textual requirements have been the object of analysis (apart from the student essay, other genres like the research article, the research report, or the grant proposal are currently analysed), and, secondly, an emphasis on the social situation of writing, understood as the impact of expectations and norms of discourse communities (cultural and disciplinary) on the shaping of discoursal practices. As it is now, Intercultural Rhetoric can be broadly defined as "a research field that seeks to identify and explain some of the rhetorical and stylistic accommodations that multilingual writers need to make in order to achieve their communication goals interculturally" (Moreno 2013, 1).

Perhaps the most relevant characteristic of this intercultural rhetoric is its "interdisciplinarity in its theoretical and methodological orientation" (Connor 2004b, 291). It draws on research methods from second language acquisition, composition and rhetoric, anthropology, translation studies, discourse analysis, corpus studies and genre analysis. In her article, Connor (2004b) evaluates three major methodological approaches for written intercultural rhetoric research: text analysis, genre analysis, and corpus analysis. The three are non-exclusive and share their sensitivity towards processes, contexts and particular situations. Moreover, an entirely new concept to the analysis of intercultural rhetoric is proposed. It is the case of oral discourse, labelled as "ethnographic approaches".

With the expansion of the research field of writing in EAP, Genre studies has been the main source of methods of analysis that has supplemented the discourse analysis frameworks used in previous contrastive rhetoric research. As Connor (2004b, 296) clearly states, "the development of genre analysis has been beneficial for intercultural rhetoric research as it has forced researchers to compare apples with apples". Corpus analysis has contributed with concepts such as equivalence or *tertium comparationis* to the design, data collection and analysis of comparable texts in intercultural studies.

The establishment of the *tertium comparationis* has been one of the most controversial issues and a fundamental methodological question in Contrastive Rhetoric. For a long time, the term of comparison was looked at from a mere structural point of view, in tune with the understanding of language as a complex hierarchical structure operating at various levels of organisation. However, written discourse is mainly a communicative activity, and, as such, it is constrained by the same basic situational factors, that is, by setting, participants, topic and purpose. As Markannen

et al. (1993, 150) suggest, "keeping these factors constant would guarantee the comparability of texts produced in those conditions", which in academic writing means comparability as far as writers are concerned (students/teachers) and comparability of genre (research articles, reviews, abstracts, dissertations, etc). The problems which might then arise in the design of any quantitative contrastive study seem to be easily solved by controlling as many of the variables as possible in the situations in which the data are collected.

It is important that we compare elements that can in fact be compared. The concept of *tertium comparationis* or common platform of comparison is essential at all levels of research: in identifying texts for corpora, in selecting textual concepts to be studied in the corpora, and in identifying linguistic features that are used to realize these concepts. It is clear from contrastive analysis that the concept of equivalence or *tertium comparationis* is a relative one and that the original idea of *identity* is giving way to the idea of *maximum similarity* (Connor and Moreno 2005).[2]

## 3. The construction of a comparable corpus: The case of the SERAC corpus (InterLAE)

Within this working framework which integrates intercultural rhetoric, genre analysis and corpus studies, I would like to illustrate the InterLAE experience in the construction of a comparable corpus and its application to intercultural research. Intercultural studies carried out by members of the InterLAE team have intended to provide answers for questions which arose in the course of our research and which are basic to any contrastive rhetoric study. Moreno (2008, 26) defines these basic questions as follows:

1. whether the imputed cross-cultural differences in the rhetorical configuration of texts actually exist,
2. if they exist, which cultural or educational factors may help to account for such differences (e.g. values, norms, learning processes and educational trends),
3. which precise difficulties with discourse structure and other rhetorical features second language users from a given non-English writing culture experience when writing in English as an L2,
4. whether difficulties experienced with discourse structure and other rhetorical features by L2 users of English are attributable to interference (or negative transference) from the first language.

The InterLAE research group has tried to provide answers to the questions above by constructing the SERAC corpus (Spanish/English Research Article Corpus), which allows comparisons in terms of:

(i) English L1 and English L2
(ii) English L1and Spanish L1
(iii) Spanish L1 and English L2
(iv) English L1, English L2 and Spanish L1

The first contrast allows to gather insights as to how Spanish writers draft their academic texts in English in comparison with native peers (L1 English writers), focusing on divergent aspects in terms of rhetorical structure, and discoursal, stylistic and lexicogrammatical aspects, providing answers to questions 1, 2 and 3. The second contrast allows comparison of rhetorical, discoursal and linguistic uses in academic texts in two languages as L1, which provides researchers with insights as to the degrees of divergence and similarity between two "big cultures" (Holliday 1999) in their rhetorical, discoursal and linguistic conventions in academic writing. This second type of contrast also provides answers to question 2 and allows to predict answers for question 3, that is, the difficulties encountered by non-native speakers when writing their texts in English. The third type of comparison may provide information and explanations for some rhetorical, discoursal and lexico-grammatical aspects of interference detected in texts written in English by Spanish academics. The last contrast allows to relate insights from the three previous types of contrast. Thus, it provides answers for question 3 which might then be complemented with answers to question 4, as it allows to attribute divergences between uses in English L1 and English L2 (and, therefore, difficulties for L2 writers) to Spanish L1 interference, these divergences giving way to a certain degree of "hybridity" in the linguistic uses of Spanish academics writing in English, that is

> the mixing of the use of the normative Standard English (pointing towards homogeneity) and the use of the local, culture-specific textual organization and textual preferences of the non-native English scholars (leading to cross-cultural heterogeneity and diversity in Academic English). (Mauranen et al. 2010, 644)

The compilation of the SERAC corpus involved taking important decisions in order to ensure cross-cultural comparability. As Pérez-Llantada states (2008, 92),

> the corpus needed to be "sufficiently representative to identify the text-linguistic preferences of Anglophone academics writing in English, and Spanish academics writing both in English and in Spanish […] and as to whether Spanish scholars transfer their culture-specific discoursal features or rather adopt the Anglophone discursive practices when writing articles in English.

Representativeness, then, ensured reliability at the time of drawing conclusions and implementing results as pedagogical materials.

The current version of SERAC (version 2.0) comprises articles written (i) in English by Anglophone scholars (English L1, subcorpus ENG), (ii) in English by Spanish scholars (English L2, subcorpus SPENG) and (iii) in Spanish by Spanish scholars (Spanish L1, subcorpus SP). To cater for cross-disciplinary contrast, SERAC represents the major academic divisions covering three disciplines each: Humanities and Arts Social Sciences, Biological and Health Sciences, and Physical Sciences and Engineering. Table 1 shows the disciplines included in each academic division, the number of texts per discipline, per subcorpus and number of words.

Two criteria were set for the selection of disciplines: the availability of the texts through the University databases and the availability of texts written in English by Spanish scholars. In fact, in certain disciplines there is limited availability of texts since the scholars either do not publish internationally in English (this seems to be the case, for instance, of Sociology, as explained in footnote 3) or use translators when writing in English. The latter is the case of many disciplines in the humanities, whereas there seems to be a tendency for hard scientists to write directly in English (Burgess et al. 2012).

As for corpus compilation, all the research articles (RAs) in the corpus were assumed to have been written in Spanish or English by expert members of the corresponding academic disciplines, published in the case of ENG and SPENG in some of the most widely-read academic journals of each field. Other factors affecting the production of texts have also been taken into account in the corpus design. One of them has been the time span for the publication of the texts. SERAC 2.0 comprises texts from 2000 to 2010. A compact time span ensures that they are representative of current academic prose. Then, the fact that, whenever possible all the ENG and SPENG RAs were selected from the same three international journals in each discipline guaranteed likely audiences, similar publication impact, a similar peer review system and editorial gate-keeping (Pérez-Llantada 2008, 95). Scholars and university researchers at the University of Zaragoza contributed with information about impact journals in their fields, also for the case of the SP journals, with no or very low impact factors.

| Academic divisions | Disciplines | ENG | SPENG | SP | Total no. of texts |
|---|---|---|---|---|---|
| Humanities and Arts | Applied Linguistics (AL) | 30 | 30 | 30 | 90 |
| | Literature (LIT) | 30 | 30 | 30 | 90 |
| | Information Science (IS) | 30 | 30 | 30 | 90 |
| No. of words | | | | | 1 946 562 |
| Biological and Health Sciences | Haematology (HAE) | 30 | 30 | 30 | 90 |
| | Urology (UR) | 30 | 30 | 30 | 90 |
| | Oncology (O) | 30 | 30 | 30 | 90 |
| No. of words | | | | | 860 694 |
| Physical Sciences and Engineering | Earth Sciences (ES) | 30 | 30 | 30 | 90 |
| | Food technology (FT) | 30 | 30 | 30 | 90 |
| | Mechanical Engineering (ME) | 30 | 30 | 30 | 90 |
| No. of words | | | | | 1 152 538 |
| Social Sciences and Education | Business Management (BM) | 30 | 30 | 30 | 90 |
| | Geography (Geo) | 30 | 30 | 30 | 90 |
| | Sociology (Soc) | 30 | 7*[4] | 30 | 67 |
| No. of words | | | | | 1 769 081 |
| Total no. of words | | 360 | 337 | 360 | 1057 |
| | | 2 146 347 | 1 771 727 | 1 811 071 | 5 729 145 |

**Table 1. SERAC 2.0. Number of texts per discipline and subcorpus. Total number of texts per discipline. Total number of texts in corpus. Number of words per academic division and per subcorpus. Total number of words in corpus.[3]**

To ensure *tertium comparationis* in the texts included in SERAC, meant, among other things, to make sure that the texts written by Spanish academics in English had not gone through language revisors, translators or any other "literacy brokers" (Lillis and Curry 2006). In this way we needed to ensure that the SPENG texts were comparable to those of the ENG subcorpus. To do this, we carried out a validation process for those academic disciplines included in SERAC where intercultural contrastive studies were going to be carried out (e.g. economics, urology, food technology, and mechanical engineering among others). Thus, a message by email was sent to the corresponding authors of the SPENG texts to ask them whether their papers had undergone any translation or revision processes and, if so, whether revision had been substantial or had only entailed minor changes in the manuscript. For corpus compilation, texts by authors who did not respond, and which had been translated or which had undergone major revision were rejected and substituted for by others of which we could ensure comparability (Pérez-Llantada 2008, 96).

## 4. Illustration of studies. Some results

The compilation of the SERAC corpus has opened a wide path of research for the InterLAE group basically across two lines, the cross-cultural and the cross-disciplinary. In this section I will refer to some representative InterLAE studies which have emerged from the discursive exploration of RAs written in by Spanish academics both in English and Spanish.

### 4.1. Intercultural analysis: L1 (English) and L1 (Spanish)

Intercultural analyses within the InterLAE research group firstly focused on the cross-cultural exploration of two genres: RAs and RA abstracts written in English and Spanish as L1.

To start with, Business Management RAs written by international scholars at Anglo-Saxon institutions were contrasted with RAs in the same discipline written in Spanish by Spanish scholars. The focus of study was the frequency of use and distribution of interactive and interactional metadiscourse features (Mur Dueñas 2011). Although previous analyses of metadiscourse (e.g. Hyland 2005) were taken into consideration during the analysis, these were modified and adjusted in the light of the data extracted from the corpus following a corpus-driven methodology to best suit the cross-cultural exploration undertaken. Thus, with regard to interactive metadiscourse, the resulting taxonomy consisted of logical

markers, code glosses, topicalisers, endophoric markers and evidentials. Interactional metadiscourse included hedges, boosters, attitude markers, engagement markers and self-mentions.

Interestingly, the analysis revealed that Spanish Business Management scholars include fewer metadiscourse features in their texts than their American-based peers. Spanish scholars in this field provide fewer explicit signals of the relationship between ideas and the organisation and clarification of ideational material in their RAs (i.e. interactive metadiscourse), as well as fewer explicit indications to their readers of their stance (i.e. interactional metadiscourse). As a consequence, readers of the Spanish RAs are provided not only with fewer features to help them navigate through the texts, but also with fewer signals indicating the authors' values and opinions than readers of the RAs in English.

On account of the findings, it could be argued that Spanish Business Management scholars addressing a local community in Spanish tend to be less overt than their peer American-based scholars addressing the international community in English in the unfolding of their texts and in their representation as authors, as members of a given disciplinary community. In all, these rhetorical differences influence the type of writer-reader relationship that is built in the two contexts.

The cross-cultural exploration of Business Management writing literacies did also include the study of RAs rhetorical structure (MurDueñas 2010), where significant differences in the rhetorical organisation of the Introduction section in RAs in the two languages were found. Moreover, a pattern of the moves and steps found in Business Management RA Introductions was outlined. A comparison with similar research in other fields reveals that some rhetorical steps seem to be specific to this disciplinary field.

The contrastive analysis of Applied Linguistics RA abstracts written in English for an international audience and in Spanish for a local audience also unearthed significant cross-cultural differences. In some of such studies (Lorés Sanz 2006; Lorés Sanz and Murillo Ornat 2007), an analysis of text-internal (lexico-grammatical) features (Bhatia 2004) was carried out, addressing the use of interactive and interactional metadiscourse (Hyland 2005) and, more specifically, of pronouns (as self mentions and engagement markers) and evidentials. Results were very much in line with the ones mentioned above for RAs in the field of Business Management before (Mur Dueñas 2007, 2011). More specifically, Spanish scholars make use of fewer metadiscoursal features in general with the exception of the inclusive pronoun *we* as an engagement marker, more frequently found in Spanish, especially in

introductory sections. It was concluded that the rhetorical purpose at work here was the need to meet the reader's expectations of inclusion with the aim of complying with interpersonal solidarity and in-group disciplinary membership. With regard to self mentions (both *I* and exclusive *we*), applied linguists publishing in English at an international level seem to make a more frequent use of exclusive pronouns, thus projecting a stronger and firmer authorial position than their Spanish peers. In contrast, writers publishing in Spanish in journals of a lower impact construct that authorial position by using the exclusive *we* even in single-authored texts, which allows them to claim authority and respect as scholars but without taking so much risk, as the use of *we* has the effect of "diluting" the authorship in a plural responsibility. With respect to the use of evidentials, in agreement with the results for the Business Management RAs (Mur Dueñas 2011) it was found that they were much more frequently used in English than in Spanish Applied Linguistics abstracts, perhaps because more pressure is felt to acknowledge previous contributions in order to claim a position in a well-established tradition.

On the whole, according to the L1 (English-Spanish) cross-cultural analyses carried out by members of the InterLAE group, different relationships between writers and their readers seem to be established at the international level in English and at the national level in Spanish. The pressures to seek recognition, to get approval and acceptance are stronger at the international level, and the competition fiercer, which results in the performance of much more prominent authorial roles which anticipates the response of rather critical readers and, on the whole, influences the writer-reader relationship established.

### 4.2. Intercultural analysis: L1 (English), L2 (English)

Cross-cultural studies have also been carried out using part of the SERAC corpus, this time comparing English L1 and L2 by Spanish academic writers. The work by Carciu (2009) illustrates this contrastive perspective in two significant pieces of research. Carciu explores the rhetorical preferences of non-native English and native English scholars in the use of the exclusive *we* pronoun and related forms (*us* and *our*) when publishing RAs in the Health Sciences internationally. Her quantitative results show that, overall, Spanish writers tend to use *we* pronouns more than their native counterparts, thus making themselves more visible in their texts particularly in Introduction and Discussion sections. However, her analysis of the discoursal functions of these pronouns following Tang and John's (1999) taxonomy indicates that they mainly have an

"exclusive-collective" referential value in both linguistic and cultural contexts, with a large number of authors engaged in collaborative work. Thus, although quantitative data point towards cross-cultural variability, genre requirements and disciplinary factors may account for the similar use of *we* pronouns in terms of their discourse roles in each RA section. This similarity may also be due to the growing interest of Spanish academics (under increasing institutional pressure) in publishing in international scientific journals, which leads them to adjust to the conventions prevailing in English-medium international scientific texts.

## 4.3. Intercultural analysis: L1 (English), L2 (English) and L1 (Spanish)

As a third step, and in the light of the differences found between the texts in English and in Spanish as L1, the possible transfer of features from Spanish L1 into English L2 has been explored. These studies have mainly been carried out in two disciplines/divisions of the SERAC corpus: Business Management, within the Social Sciences division, and the Biomedical and Health Sciences division. In all cases, the corpora used for the study included RAs written in English L1, by scholars based at Anglo-Saxon institutions, in Spanish L1 and in English L2 also by Spanish scholars. A major result from these analyses is that there are some discoursal and rhetorical conventions which are more culturally-engrained and thus more likely to be transferred than others.

Taking the discipline of Business Management as the field of analysis, Murillo Ornat (2012a) focuses on the exploration of reformulation markers in English L1, English L2 and Spanish as L1. The general frequency of use of the markers, the types of markers used, the functions most commonly performed and their (non-)parenthetical uses are compared in order to explore the degree of transference in their use by the L1 Spanish scholars writing L2 English RAs. The results lead to conclude that some general rhetorical Spanish L1 features are more likely to be adapted in the L2 English texts written by L1 Spanish academics than other more specific grammatical features.

Logical and reformulation markers (additive, contrastive and consecutive) are also explored in the discipline of Mechanical Engineering in English L1 and L2 and in Spanish L1, and the results compared with previous studies on Business Management (Mur Dueñas 2009; Murillo Ornat 2012b). While there are no statistically significant differences between the distribution of the markers in the English and Spanish-English subcorpora, there are differences in the specific logical and reformulation

markers used. We can thus perceive some signs of the existence of an intermediate code between the languages involved. The study of logical and reformulation markers carried out in the Business Management subcorpus did not yield any distributional differences between the two English subcorpora. From this we may infer that the processes of transference and adjustment between languages seem to be different depending on the disciplines.

The authorial voice by Spanish Business Management scholars in English has also been explored contrasting English L1, L2 and Spanish L1 (Lorés Sanz 2011a). Results have shown divergences in the frequency and distribution of the sequence under investigation (*we*+verb), both in terms of function and tense (including the association with modal verbs), which reveal the existence of pragmatic and cultural factors that may hinder the projection of a firm, confident authorial voice by Spanish academics in an increasingly competitive international academic environment.

The construction of the authors' voice has also been investigated in the frequency of use, distribution, and discourse function of first person pronouns across English and Spanish in Business Management RAs (Lorés Sanz 2011b). The divergences observed suggested that the disciplinary and the linguistic and cultural variables interplay and determine the textual points in the research article at which authors make themselves visible, and the frequency of that visibility. Significant conclusions can then be drawn as to how visible Spanish academics make themselves when writing in English for an international audience, and whether the divergences/similarities found in comparison with their Anglo-American peers respond to interference of their academic literacies in Spanish. This knowledge may then allow them to make informed decisions as to whether and/or in what senses they should modulate their voice to comply with what is expected from writers publishing in international contexts in the field of Business Management.

As mentioned above, the Biomedical and Health Sciences disciplinary division has also been object of exploration from an intercultural perspective. Here, the studies by Pérez-Llantada (2010a, 2010b) and Carciu (2013) should be mentioned. Adopting Ädel's (2006, 2008) taxonomy of text-oriented and participant-oriented functions of metadiscourse, Pérez-Llantada (2010a) carries an exploration of metadiscourse in RAs in the Health Sciences written by scholars from two cultural contexts (North-American and Spanish) and in two languages (English and Spanish) in order to identify the micro-level discourse functions of metadiscourse in Introductions and Discussions in the texts under study. She also explores the correlation between these functions and

the information-organising moves established for these sections. Her results show similar uses of micro-level discourse functions in each rhetorical section in the three subcorpora but also both culture- and language-specific lexico-grammatical realisations of metadiscourse units. These findings point towards a process of accommodation to the patterns prevailing in international publications which seems to be taken place among non-native (Spanish) scholars, and which can be taken to be a clear indicator of the gradual homogenisation and standardisation of writing processes in academic English (cf. Mauranen et al. 2010).

A study of epistemic lexical verbs is also carried out in biomedical RAs in the three linguistic and cultural environments (Pérez-Llantada 2010b). Her study shows that the use of epistemic modality tends to be highly routinised in the three textual environments and that the Spanish writers writing in English modalise their discourse with epistemic meanings as the Anglophone scholars do. This considerable degree of homogeneity suggests a possible effect of globalization affecting the writing practices in the two cultural contexts. However, as the author points out, the hybrid nature of the English L2 texts in certain linguistic aspects points towards current concerns about the possible effects of the academic and research globalization process on the production of non-native English writers publishing internationally.

Carciu (2013) explores the phraseological profile of *we* first-person pronoun references in native and non-native biomedical research article Introduction sections and addresses the question of whether culture-specific features can be explored through phraseological items in written academic English. Her results reveal, for instance, that Spanish scholars writing in Spanish show a high degree of impersonality, almost avoiding the use of n-grams with *we* references, whereas when they write in English L2 they display a much greater degree of formulaicity. In Carciu's view whereas similarities between English L1 and English L2 can be explained by the disciplinary and genre-specific phraseology, differences can be interpreted as a sign of the L1 vs L2 language user status. In all, the research article can be taken to be a negotiated intercultural space which promotes a shared disciplinary identity across cultures. However, her results also indicate that the linguistic expression of identity throughout the research article does not completely erase cultural identities.

## 5. Applications of intercultural research to EAP teaching

The development of pedagogical applications resulting from research in academic writing is one of the main and most obvious objectives when

the intercultural perspective is adopted. In Spain, studies such as the ones carried out by Fernández Polo and Cal Varela (2009), at the University of Santiago de Compostela, or Ferguson, Pérez-Llantada and Plo (2011), at the University of Zaragoza, reveal the positive attitude that Spanish academics have towards English as a language of international communication, but they also show that there are great differences in the use of EAP among different areas and disciplines, and the strong training needs that Spanish researchers, especially in certain fields, have.

Another, more ambitious, study along the same lines is the one put forward by the ENEIDA group (*Equipo Nacional de Estudios Interculturales sobre el Discurso Académico*), which, unlike other research based on a single institution, takes into account the data of five main Spanish institutions: the Spanish National Research Centre (CSIC), Universidad de León, Universidad de La Laguna, Universitat Jaume I and Universidad de Zaragoza, all of them long-established and prestigious universities. The answers of over 1,700 informants to a very detailed questionnaire point towards significant differences in terms of disciplinary community in the level of competence of Spanish academics in EAP, in how they perceive their needs as members of a disciplinary community which projects itself internationally, and, more importantly, in their training needs, revealing that Spanish academics are at different stages of an "internationalization" process.

The design of materials and implementation of courses in academic writing is one of the main objectives of the InterLAE research group. One of such pedagogical applications was a course consisting in a series of workshops addressed to and tailored-made for teachers and researchers at the Faculty of Arts in the University of Zaragoza.[5] The rationale for those workshops were the areas of linguistic, cultural, disciplinary and/or generic difficulty for Spanish academics writing in English that we had been able to detect in our previous research. There were five workshops, whose contents and methods were designed and specified attending to the disciplinary and generic needs, lacks and wants of the participants. Valuable information for the design of the course was previously gathered through a questionnaire about their research interests and needs, following an ESP genre-based writing instruction approach (Swales 1990, 2004; Bhatia 1993; Flowerdew 1993, 2002; Swales and Feak 2000, 2004; Hyland 2003, among others). The five workshops covered the following topics: general aspects of academic writing, the writing of abstracts, the writing of research articles, correspondence and emailing in the academia, and bionotes and CVs.

Three main groups of scholars attended the workshops: linguists, historians and geographers. To cater for evident disciplinary differences, materials had to be created and adapted attending to the scholars' research backgrounds. Along the five workshops special attention was paid to cross-cultural differences between academic literacies in Spanish and in English at all levels, from the microlevel of lexico-grammatical features to a more discoursal perspective, focusing on aspects such as the use of hedges, modal verbs, boosters and personal pronouns.

By implementing this course, we were able to show that the insights and conclusions we had gathered from our own previous contrastive corpus studies could be applied and shaped into instructional materials.

## 6. Limitations

As I have tried to illustrate in this chapter, in our view cross-culturality is best explored in academic writing by means of corpus studies, perhaps the only way of drawing significant insights as to the impact of cultural and linguistic identity on the non-native's written production in English.

However, several issues have arisen in the course of our research that deserve to be mentioned here. One of them is the use of published RAs for our studies. The question is whether these texts can help us explore difficulties and challenges for Spanish academics when they have already been approved of by gatekeepers. To cater for this limitation, studies have started to be carried out on what Lillis and Curry (2006, 2010) call "text histories", that is, the collection of texts, from the very first draft to the final version (published or rejected), where we can trace which aspects are being commented on, suggested change or corrected by wordface professionals and gatekeepers.This is the case for instance of Mur Dueñas (2012b), which focuses on the text histories of a team of Spanish researchers in the field of Finance who struggle to get their research articles published internationally in English. There may be a crossing line beyond which publication may be hindered if international conventions are not partially accommodated. The question is to know where this crossing line lies and whether it changes over time. A closer look into the writing process, not just the products, may allow us to gain a better insight into it.

A second relevant issue is the question of English used as a lingua franca. All the texts included in the English L1 subcorpus in SERAC 2.0. were written by academics affiliated to Anglophone institutions. Thus, our research takes English as used by Anglophone speakers as a parameter of comparison, and not English as an international language of scientific communication. This has to be borne in mind when drawing conclusions

as to the differences between English L1 (ENG) and English L2 (SPENG) found in our corpus. However, as Mauranen et al. (2010) claim, if our aim is to explore present-day academic language in English, we should look at the way English works as a lingua franca (ELF) and not at English as a native language, because ELF is a better representative if we want to capture the nature and features of the academic English used in international publications (2010, 640):

> Strictly speaking, academic discourses in themselves have not native speakers: they are learned in secondary socialization by all participants in academic communities of practice. Issues of register, specific terminology and phraseology, along with mastery of relevant genres, acceptable modes of argumentation, and ways of presenting a case are all consciously learned skills which are not acquired in the same way as a mother tongue.

To cater for this new dimension, a fourth subcorpus of ELF RAs from all the disciplines included in the corpus is planned to be incorporated to SERAC. This new collection of texts will cater for contrastive research across different linguistic and cultural contexts, allowing comparisons, for instance, between academic writing in ELF in general vs ELF as written by Spanish academics, and/or English L1 vs ELF. Results from these studies will hopefully provide further insights into the difficulties and conflicts of Spanish scholars' linguistic, discoursal and rhetorical choices when using English to communicate in our globalized scientific world.

## Acknowledgement

I would like to thank my colleagues Pilar Mur, Silvia Murillo, Oana Carciu and Carmen Pérez-Llantada for the helpful information they provided to write sections 3 and 4 of this chapter.

## Notes

[1] This research has been carried out within the framework of the research group InterLAE (www.interlae.com), financially supported by the Spanish Ministerio de Economía y Competitividad with the project "English as a lingua franca across specialised discourses: a critical genre analysis of alternative spaces of linguistic and cultural production" (FF12012-37346).

[2] Moreno (2008) offers a model of how this requirement of maximum similarity could be met in inter-cultural studies that draw on one particular theoretical framework, genre theory. In her model, she describes a series of similarity constraints which should be taken into account to construct a *tertium comparationis* to carry out inter-cultural analysis. These similarity constraints

include, for instance, text form, genre, participants, academic discipline and general purpose of communication, among others.

[3] The low number of texts in the SPENGSOC corpus is due to the impossibility to find texts that comply with the comparability criteria applied to the rest of the SERAC corpus, that is, texts written in English by Spanish academics and published in prestigious international publications in the last ten years.

[4] Data extracted from Pérez-Llantada (2012, 73-77).

[5] A full report of this pedagogical experience is included in Mur and Lorés Sanz (2010).

## References

Ädel, Annelie. 2008. Metadiscourse across three varieties of English. American, British and Advanced-learner English. In *Contrastive rhetoric: Reaching to intercultural rhetoric*, ed. Ulla Connor, Ed Nagelhout and William Rozycki, 45-62. Amsterdam/Philadelphia: John Benjamins.

—. 2006. *Metadiscourse in L1 and L2*. Amsterdam/Philadelphia: John Benjamins.

Atkinson, Dwight. 2004. Contrasting rhetorics/contrasting cultures: Why contrastive rhetoric needs a better conceptualization of culture. *Journal of English for Academic Purposes* 3(4): 277-289.

Bhatia, Vijay K. 1993. *Analyzing genres: Language use in professional settings.* London: Longman.

—. 2004. *Worlds of written discourse: A genre-based view.* London: Continuum.

Burgess, Sally, Pilar Mur Dueñas, Rosa Lorés Sanz, Jesús Rey Rocha, and Ana. I. Moreno. 2012. Underthreat from all sides: Historians and English for research publication. Paper presented at the conference *English in Europe: Debates and Discourses*. Sheffield, UK, 20-22 April, 2012.

Carciu, Oana. 2009. An intercultural study of first-person plural references in biomedical writing. *Ibérica* 18:71 -92.

—. 2013. Formulating identity in academic writing across cultures: N-grams in introduction sections. *ESP across Cultures* 10: 87-109.

Connor, Ulla. 2004a. Introduction to *Journal of English for Academic Purposes* 3(4): 271-276.

—. 2004b. Intercultural rhetoric research: Beyond texts. *Journal of English for Academic Purposes* 3(4): 291-304.

Connor, Ulla, Ed Nagelhout, and William Rozycki, eds. 2008. *Contrastive rhetoric: Reaching to intercultural rhetoric*. Amsterdam: John Benjamins.

Connor, Ulla, and Ana I. Moreno. 2005. Tertium comparationis: A vital component in contrastive research methodology. In *Directions in applied linguistics: Essays in honor of Robert B. Kaplan*, ed. Paul Bruthiaux, Dwight Atkinson, William G. Eggington, William Grabe and Vaidehi Ramanathan, 153-164. Clevedon, England: Multilingual Matters.

Ferguson, Gibson, Carmen Pérez-Llantada, and Ramón Plo. 2011. English as an international language of scientific publication: A study of attitudes. *World Englishes* 30:41-59.

Fernández Polo, Javier, and Mario Cal Varela. 2009. English for research purposes at the University of Santiago de Compostela: A survey. *Journal of English for Academic Purposes* 8: 152-164.

Flowerdew, John. 1993. An educational, or process, approach to the teaching of professional genres. *ELT Journa*l 47: 305-316.

—. 2002. Genre in the classroom: A linguistic approach. In *Genre in the classroom: Multiple perspectives*, ed. Ann Johns, 89-100. Mahwah, NJ: Lawrence Erlbaum.

Holliday, Adrian. 1999. Small cultures. *Applied Linguistics* 20(2): 237-264.

Hyland, Ken. 2003. *Second language writing.* Cambridge: Cambridge University Press.

—. 2005. *Metadiscourse*. London: Continuum.

Kaplan, Robert K. 1966. Cultural thought patterns in inter-cultural education. *Language Learning* 16(1-2): 1-20.

—. 1988. Contrastive rhetoric and second language learning: Notes toward a theory of contrastive rhetoric. In *Writing across languages and cultures*, ed. Alan C. Purves, 275-304. Newbury Park: Sage Publications.

Lillis, Theresa M., and Mary Jane Curry. 2006. Professional academic writing by multilingual scholars: Interactions with literacy brokers in the production of English-medium texts. *Written Communication* 23(1): 3-35.

—. 2010. *Academic writing in a global context*. London: Routledge.

Lorés Sanz, Rosa. 2006. "I will argue that": First person pronouns and metadiscoursal devices in RA abstracts in English and Spanish. *ESP across Cultures* 3: 23-40.

—. 2011a. The study of authorial voice: Using a Spanish-English corpus to explore linguistic transference. *Corpora* 6(1): 1-24.

—. 2011b. The construction of the author's voice in academic writing: The interplay of cultural and disciplinary factors. *Text & Talk* 31(2): 173-193.

Lorés Sanz, Rosa, and Silvia Murillo Ornat. 2007. Authorial identity and reader involvement in academic writing: A contrastive study of the use of pronouns in RA abstracts. In *Aprendizaje de lenguas, uso del lenguaje y modelación cognitiva: Perspectivas aplicadas entre disciplinas,* ed. Ricardo Mairal et al., 1249-1257. Madrid: UNED.

Markkanen, Raija, Margaret S. Steffensen, and Avon Crismore. 1993. Quantitative contrastive study of metadiscourse problems in design and analysis of data. *Papers and Studies in Contrastive Linguistics* 23: 137-151.

Mauranen, Anna. 1993. Contrastive ESP rhetoric: Metatext in Finnish-English economics texts. *English for Specific Purposes* 12: 3-22.

Mauranen, Anna, Carmen Pérez-Llantada, and John M. Swales, 2010. Academic Englishes: A standardised knowledge? In *The world Englishes handbook*, ed. Andy Kirkpatrick, 634-652. London, New York: Routledge.

Moreno, Ana I. 2008. The importance of comparable corpora in cross-cultural studies. In *Contrastive rhetoric. Reaching to intercultural rhetoric,* ed. Ulla Connor, Ed Nagelhout and William Rozycki, 25-41. Amsterdam: John Benjamins.

—. 2013. Intercultural rhetoric in language for specific purposes. In *The encyclopedia of applied linguistics*, ed. Carol A. Chapelle, 1-5. Oxford: Blackwell Publishing.

Mur Dueñas, Pilar. 2007. "I/we focus on…": A cross-cultural analysis of self mentions in business management research articles. *Journal of English for Academic Purposes* 6(2): 143-162.

—. 2009. Logical markers in L1 (Spanish and English) and L2 (English) business research articles. *English Text Construction* 2(2): 246-264.

—. 2010. A contrastive analysis of research article introductions in English and Spanish. *Revista Canaria de Estudios Ingleses* 61: 119-133.

—. 2011. An intercultural analysis of metadiscourse features in research articles written in English and in Spanish. *Journal of Pragmatics* 43(12): 3068-3079.

—. 2012. Getting research published internationally in English: An ethnographic account of a team of finance Spanish scholars' struggles. *Ibérica* 24:139-156.

Mur Dueñas, Pilar, and Rosa Lorés Sanz. 2010. Responding to Spanish academics' needs to write in English: From research to the implementation of academic writing workshops. In *Ways and modes of human communication*, ed. Rosario Caballero Rodríguez and María

Jesús Pinar Sanz, 501-510. Cuenca (Spain): AESLA y Ediciones de la Universidad de Castilla-La Mancha.

Murillo Ornat, Silvia. 2012a. The use of reformulation markers in business management research articles: An intercultural analysis. *International Journal of Corpus Linguistics* 17(1): 62-88.

—. 2012b. Discursive insights for a cross-cultural analysis: Logical and reformulation markers in L1 English and Spanish and L2 English mechanical engineering research articles. Paper presented at the *2nd International PRISEAL Conference*: *Publishing Research Internationally: Issues for Speakers of English as an Additional Language*. University of Silesia, Sosnowiec/Katowice (Poland), 9-11 June 2011.

Pérez-Llantada, Carmen. 2008. Humans vs machines? A multi-perspective model for ESP discourse analysis. *ESP across Cultures* 5: 91-104.

—. 2010a. The discourse functions of metadiscourse in published writing. Issues of culture and language. *Nordic Journal of English Studies* 9(2): 41-68.

—. 2010b. The "dialectics of change" as a facet of globalisation: Epistemic modality in academic writing. In *English for professional and academic purposes*, ed. Miguel F. Ruiz-Garrido, Juan Carlos Palmer-Silveira and Inmaculada Fortanet-Gómez, 25-42. Amsterdam, New York: Rodopi.

—. 2012. *Scientific discourse and the rhetoric of globalization. The impact of culture and language*. London and New York: Continuum.

Purves, Alan. 1988. Introduction to *Writing across languages and cultures,* ed. Alan Purves, 9-21. Newbury Park: Sage Publications.

Swales, John. 1990. *Genre analysis: English in academic and research settings*. Cambridge: Cambridge University Press.

—. 2004. *Research genres: Exploration and applications.* Cambridge: Cambridge University Press.

Swales, John, and Christine Feak. 2000. *English in today's research world: A writing guide.* Michigan: Universityof Michigan Press.

—. 2004. *Academic writing for graduate students: A course for non-native speakers of English.* Michigan: University of Michigan Press.

Tang, Ramona, and Suganthi John. 1999. The "I" in identity: Exploring writer identity in student academic writing through the first person pronoun. *English for Specific Purposes* 18(Supplement 1): S23-S39.

# PART III

# ENGLISH AS LINGUA FRANCA IN ACADEMIC SETTINGS

# CHAPTER EIGHT

# 'INTERNATIONALITY' AS A METAPRAGMATIC RESOURCE IN RESEARCH PRESENTATIONS ADDRESSED TO *ENGLISH AS A LINGUA FRANCA* AUDIENCES

LAURIE ANDERSON
UNIVERSITÀ DI SIENA, ITALY

## 1. Introduction: The changing face of today's academia

In recent volumes Jenkins (2007, 2014), Seidlhofer (2011), Mauranen (2012) and others have highlighted the importance of reconceptualising ELF (English as a Lingua Franca) communication in terms that take into account the radical transformation of social relations that has occurred as a result of globalization. Seidlhofer observes that "relations, transactions, and networks have […] become much more extensive and cut across conventional communal boundaries, transforming the very concept of community in the process" (2011, 83). Mauranen (2012) concurs, observing that among the characteristics that distinguish ELF communication from communication in traditional speech communities are "non-locality, non-permanence, speaker mobility and multilingualism" (Mauranen 2012, 23).

Nowhere is the impact of these changes more evident than in the domain of higher education. Student mobility, both within and across regions, is on the rise. In Europe, for instance, the face of university classrooms has changed considerably in the last two decades thanks to EU mobility programs that have increasingly brought students from different countries into contact face-to-face; in recent years these students have been joined by others from both BRIC (Brazil, Russia, India, China) and developing countries who choose to go abroad for an education they hope will give them a competitive edge. The presence of increasingly culturally-

diverse student populations in universities in Europe and North America is not the only manifestation of globalization in higher education: another is the establishment by Western universities of 'off-shore satellites', such as the Georgetown campus in Doha (Qatar) or NYU's campus in Abu Dhabi (U.A.E.); yet another, the emergence of MOOC (Massive Online Open Courses) and other electronically-delivered forms of instruction which challenge the very notion that pedagogic interaction requires the physical co-presence of teachers and learners. Globalizing trends are also evident in research collaboration, which is on the increase both across borders and across continents, despite the fact that currently the only political institution actively mediating relationships on this level is the European Commission (Leydesdorff and Wagner 2008). A number of factors appear to be fuelling this trend: these include both changes within academia itself, such as increasing disciplinary differentiation, and broader societal factors, such as increases in international trade and the growth of information and communications technologies (see Wagner and Leydesdorff 2005 for a more extended discussion).

While university teachers and lecturers have traditionally maintained international contacts (cf. Charle et al. 2004, for a historical overview), the developments outlined above have radically affected the professional lives of today's academics. Older established scholars are increasingly facing the challenge of teaching students of different cultural and linguistic backgrounds, not only in their national language but also in English as an academic lingua franca (Björkman 2011; Jenkins 2013). Early-career scholars are even more directly implicated, as the changing structure of academic careers and decreasing availability of tenure-track positions in many countries increasingly make mobility a requirement for career advancement (Marimon et al. 2009; Kim 2010). Marginson (2008) has argued that the current trend towards internationalization and globalisation in higher education is creating two 'tiers' of academics: on the one hand, scholars whose professional activities and horizons remain confined within national boundaries and systems; on the other, a 'coming community' of young scholars (the provocative term is borrowed from Aalbers and Rossi, 2007), whose educational profiles and professional trajectories are transnational/international in scope and often itinerant in practice.[1]

What all this means in practical terms for individual academics is only beginning to be explored, but one thing is certain: the traditional reference points for and means (both material and discursive) of establishing and maintaining professional identities have profoundly changed. When as researchers on EAP we investigate how English is used in classroom and

research settings around the globe, it is thus not just the syntax and lexis that should concern us, but also (and, indeed, perhaps primarily) the ways in which the language is being used to create and sustain interaction in this increasingly globalized 'academic space'.

The present study aims to contribute to an understanding of the pragmatics of academic ELF communication by examining the role that the thematization of self and other identities in terms pertinent to membership in an *international* community of scholars plays in peer-to-peer interaction among academics from different national backgrounds. The data examined consist of 183 research presentations by early-career scholars collected in an English-as-a-lingua-franca setting: the Max Weber Programme, an EU-funded postdoctoral programme located at the European University Institute, Florence. The program in question is designed to foster interdisciplinary research among and career advancement by early-career scholars working within the European context. Highly selective (the success rate among applicants is circa 4%), every year the programme hosts from 40 to 50 young researchers in the fields of history, economics, law and political and social sciences. The participants in the present study belong to four cohorts of fellowship holders and hail from a wide number of European and extra-European countries (up to 25 different countries, depending on the cohort). The context and data can thus be considered as highly representative of lingua franca interaction in today's globalizing academia.

## 2. Theoretical framework: Membership categorization as a metapragmatic resource

Increasing interest has been shown over the last few years in how identities are discursively established and maintained through talk. Among the various perspectives on the topic, the present analysis draws in particular on the notion of Membership Categorization (MC), an approach to the analysis of social interaction initially developed by Sacks (1972a, 1972b) and other conversational analysts working within an ethnomethodological framework. In a nutshell, MC views categories such as gender, ethnicity, national origin, occupation, and disciplinary affiliation – as well as other more 'ad hoc' categories (e.g. 'hotrodder' or 'computer nerd') – as socially relevant to the extent that they are invoked in and through talk. Such invoking can be explicit, as in the case of self or other-labelling, or implicit, as in the case of descriptions which indirectly index membership in a given category. In the following excerpt, Schegloff (2007, 471) provides a practical illustration of how using categories that

refer to discourse participants contributes to the way people participate in ongoing interaction (the illustration is, *inter alia*, pertinent to the specific focus of the present study):

> Having introduced me to one person at our first meeting as 'a sociologist,' others are readily oriented to *disciplinary* categories, and the relevance of doing so is given by the prior bit of identifying. For some next person to be identified or to self-identify as 'Canadian' is then registerable as a 'departure' from the relevancies already introduced, and can prompt a search for what has occasioned that categorization ('why that now').

A key concept in MC is the notion of so-called 'relational pairs' (Sacks 1972a).[2] The basic intuition behind this idea is that mentioning directly or referring indirectly to a certain social category (for example, 'baby') may trigger other, related categories (such as 'mother' or 'father') that then become relevant to the ongoing interaction. To take Schegloff's example, introducing one person as 'a sociologist' activates disciplinary membership as a relevant 'categorization device', thus making it pertinent for other participants at a meeting to introduce or refer to themselves or others as 'a linguist' or 'an economist'; referring to oneself or to another as 'Canadian', instead, may make relevant the geopolitical category of 'citizen of a particular nation-state'. The theoretical framework within which MC is situated is thus imminently micro-sociological; it is also constructivist, in the sense that language use is viewed as contributing to the local constitution of social order.

Membership categorization analysis has attracted the attention of scholars in various fields interested in the analysis of social interaction, from social psychology to narrative inquiry (cf. Antaki and Widdecombe 1998; Benwell and Stokoe 2006). In recent years the concept has been fruitfully employed by researchers on intercultural communication, who have used it to describe how cultural categories and national identities can be made relevant (or non-relevant) in conversations among people from different backgrounds (Day 1994; Nishizaka 1999; Zimmerman 2007). More recently, MC has also been employed by researchers working on business communication; two studies belonging to this line of inquiry are of particular relevance to the present analysis for methodological reasons.

The first is a recent article by Hougaard (2008) which details the sequential procedures through which representatives of European companies conversationally situate themselves and their respective firms as 'international' in the course of cross-borders business calls. These procedures include the listing of several different countries as sites of business and the use of geographical references that, by virtue of being

'recognitional' (rather than descriptive), cast speakers' conversational partners as 'co-Europeans' in the know. Hougaard (2008, 308) claims that such sequential practices help to characterize the speaker and his/her company as familiar with and active on the international business scene, and refers suggestively to their deployment as "doing being international". She supports this claim through several extended analyses illustrating how describing and labelling one's company and its activities in geographical and geopolitical terms plays a key role in broader narrative strategies of identity construction.

A second study of particular interest is Van De Mieroop's analysis (2008) of speeches made by business professionals to other members of the business community. The study focuses on how speakers use the audience as a discursive resource for the construction of their professional identity. Drawing on Sacks' notion of 'relational pairs' (Sacks 1972a), Van De Mieroop shows how speakers engage in 'altercasting', a practice she defines (following Weinstein and Deutschberger 1963) as "projecting an identity, to be assumed by other(s) with whom one is in interaction, which is congruent with one's own goals" (2008, 492). The author illustrates how speakers in her data project the role of potential buyer of a product or service onto the audience, in order to then present their own company in the complementary role of seller. The study is relevant to the present analysis for two reasons: first, because it focuses on a speech event similar to the one examined here (i.e. professional speeches); second, because it highlights the fact that the incumbents of categories in relational pairs can be groups of people rather than only individuals.

Drawing on the above and related studies, the present study approaches 'internationality' in academic settings as an identity construct that is achieved interactionally by participants through evoking socially-relevant categories in the course of talk. In particular, it focuses on how, in the course of research presentations addressed to academic peers, early-career scholars situate themselves and their listeners in the virtual space of international academia through reference (explicit and implicit) to *geopolitically-relevant categories* – i.e. by referring to geographical locations (e.g. countries, cities, regions), nationality, language and/or ethnicity. The aim is to show how self and other-categorization in such terms help to characterize the participants in such events as '*international* scholars'.

## 3. Data and methodology

The 183 presentations analysed for the present study took place at the beginning of each of the four one-year fellowship periods. They were addressed to an interdisciplinary audience consisting of the other fellows in the cohort and, on a rotating basis, the fellows' departmental mentors. The presentations ranged from 12 to 15 minutes in length and were followed by a brief question-answer session. The primary aim of the presentations was to familiarize the other members of the cohort with the speakers' academic background and research plans in order to encourage collaboration within and across disciplines. All of the presentations were videotaped for research and training purposes, with the speakers receiving detailed feedback as part of the academic communications component of the post-doctoral program.

An initial examination of the recordings revealed that the majority of self-references in geopolitical terms occurred in the opening sections. These sections were thus transcribed for all speakers, together with all other portions of the recordings containing self or other-categorizations in geopolitically-relevant terms. The references identified were subsequently coded by two researchers in both functional (self/other categorization) and syntactic/semantic terms (e.g. 'I'm + *nationality*') and inserted into an Excel file for easier reference and quantitative analysis.[3] A number of extended narrative sequences containing self- and other-categorizations and descriptions were also transcribed and scrutinized for recurrent patterns, thus providing both a quantitative and qualitative perspective on the data.

In what follows (section 3) I report the results of this investigation, focusing, first, on practices of self-categorization and self-description and, secondly, on practices of 'altercasting', in which speakers construe the audience in various ways as a member of a relational pair. Some brief remarks about the theoretical and applied implications of the analysis conclude the study (section 4).

## 4. Results: 'Doing being international' in academic presentations

This section is articulated into four parts. Section 3.1 presents an analysis of explicit self-categorization in geopolitical terms which highlights how, alongside straightforward statements about national origin, the speakers in my sample use a range of other types of geopolitical self-attributions. The following section (3.2) illustrates how speakers utilize

these alternative modes of self-categorization and related discourse choices to attribute transnational status to themselves through narratives about their personal and academic backgrounds. How speakers construe their audience (and reflexively, themselves) as members of a group of international scholars is the focus of the following two sections: section 3.3 shows how speakers attribute stereotypical expectations about pronunciation, spelling and other language-related issues to their listeners in order to position themselves as interculturally savvy individuals who are 'in the know' about the expectations of the international scholarly community; section 3.4 illustrates how, by construing their listeners as incumbents of relational pairs pertinent to international academia (e.g. as a group of 'European' scholars), speakers are able to characterize themselves as particular types of scholars (e.g. 'North American').

## 4.1. Self-categorization in geopolitical terms

Self-categorization in geopolitical terms appears in the data mainly in two basic types of constructions: *labelling* ("I'm...", "I'm a...") and *descriptions of origin* ("I'm from...", "I'm originally from..."; "I come from...", "I was born in..." or similar). There are also a certain number of 'WE + *ethnic or national label*' constructions (e.g. "in Hungary we...", "for us Ukrainians...").

Analyzing these utterances in terms of the geopolitical groupings they index, it is possible to identify three distinct modes of self-categorization:

1) straightforward self-categorizations in terms of *nationality* or *country of origin* (group 1);
2) self-categorizations displaying a *'local' identity* (group 2);
3) self-categorizations indexing *transnational status* (group 3).

In order to highlight this three-way division in the data, the utterances produced by all speakers who used explicit geopolitical self-categorization are reproduced below. Where a speaker produced more than one such categorization, all of the utterances produced appear together under the geopolitical category to which the first utterance belongs.[4] Within each category, the examples are further subdivided according to the specific syntactic construction used.

**Self-categorization in terms of nationality/nation-state**

*- labelling: nationality*

**(1)**
I'm Russian. (B-25)

**(2)**
I'm a Georgian national. (B-01)

**(3)**
I'm Italian and I'm a historian. (A-27)

**(4)**
As probably most of you know, I'm British. (D-22)

**(5)**
For those people who I haven't yet managed to meet, I'm English. (C-46)

**(6)**
Well, I'm Ukrainian, as you can- some of you can probably tell from my name. (B-28)

**(7)**
I'm French, I'll do my best to hide it. (A-03)

**(8)**
As you can hear I'm French [...] to be really honest, actually, I'm born in Paris. (B-40)

**(9)**
I guess I don't need to mention that I'm French, because my accent speaks for itself. Some of you already know that I come from Grenoble. (B-11)

**(10)**
(*after comment on pronunciation of name*) So I'm French. I come from Paris. (C-07)

**(11)**
(*later in presentation*) Probably for people from the European Union it's much easier but for us Ukrainians and for us, for people from other countries it's not, it's really a challenging exercise. (A-29)

**(12)**
In Hungary we don't have policy studies. (C-25)

**(13)**
I lived there (*points to map on PP*) because we did colonize uhm: this country of North Africa. However I’m an (awkward) Italian, and I knew nothing about Libya until I went to Princeton for my masters. (C-11)

**(14)**
(*later in presentation, after PP showing cover of his recent book*) Not only to give you sort of a demonstration of US-style self-promotion, which is something we have been specialising in – Europeans don’t do quite as much of that – but because some of you are interested in memory, in the role of ideas and culture, and perhaps in the Nazi past specifically. (B-03)

*- descriptions of origin: nation-state*

**(15)**
I’m from Turkey. (C-03)

**(16)**
I’m from Singapore and I’m a political scientist. (C-09)

**(17)**
I’m [NAME] and I think most of you know that I’m from Germany and I’m a lawyer, one of the many we have in our group. (A-16)

**(18)**
And I’m from from Slovenia (*laughter, photos on PP*), the most beautiful European country. And my hometown’s Ljubljana (C-29)

**(19)**
I come from Poland. (A-32)

**(20)**
I come from Thailand. (B-35)

**(21)**
I come from Greece. (C-22)

**(22)**
Well, I come from Russia, from Moscow. (A-40)

**(23)**
(*After commenting about pronunciation of name*) Okay, so it’s clear I come from China. (B-15)

**(24)**
I was born and raised in Japan. (D-36)

**(25)**
I hail from Canada and and specifically the very cold part of Canada. (B-42)

**self-categorization in terms of a 'local' identity**

*- labelling: sub-national*

**(26)**
I'm a genuine east coaster. I grew up in Atlanta, Georgia, and the humidity today is reminding me of my hometown. (A-12)

**(27)**
I'm I'm an Essex girl (*laughter*), and that means- in political science terms that means that I carry a heavy tradition of behaviouralism and qualitative methods. (C-22)

*- descriptions of origin: city*

**(28)**
I'm from Dublin. (A-10)

**(29)**
I'm from Montreal Canada. (A-14)

**(30)**
I'm from Krakow, it's a city in Southern Poland. (B-05)

**(31)**
Okay I'm from a small town in Ohio. (D-29)

**(32)**
(*later in presentation, in introducing current research*) I chose to study the case of the Grand ducat of Tuscany, not only because I'm from Florence so I'm supposed to know a little bit the archives here. (A-35)

**(33)**
My name is [NAME-SURNAME] from Buenos Aires, Argentina. (D-02)

**(34)**
I come from the most beautiful city in the world, Strasbourg. (D-43)

**(35)**
(*later in presentation, in introducing PhD research*) There’s been a huge debate in the United States and in Boston, where I come from. (C-17)

**(36)**
So I was born and raised here (*points to photo on PP*) in Gulfport Mississippi, it’s a small town near New Orleans Louisiana in the States. (D-27)

**(37)**
I was born in Rome in 1978. (C-32)

**(38)**
So I’m born in Hamburg, which is my hometown, my favourite town. (C-44)

**(39)**
So I was born in Hamburg in Germany. (B-43)

**(40)**
I was born in Berlin and did my master in Berlin, and received my PhD or will receive my PhD I hope from Humboldt University in Berlin. (B-18)

## Self-categorization in terms of transnational status

*- labelling: mixed/dual nationality*

**(41)**
I’m mostly Dutch and my academic background is a little peripatetic I guess. (D-40)

**(42)**
I’m half Iranian half Belgian and I hold French citizenship. [...] (*towards end of presentation, in discussing research objectives*) I want to use my background as a half-Iranian half-Belgian person to build bridges between civilisations. (A-29)

**(43)**
First, I’ll say that I’m Italian-Brazilian, which is a win-win situation because whenever there’s a world cup there’s a high probability that I’ll be celebrating at the end of the final. (*audience laughter*) (A-25)

**(44)**
I’m German by birth and passport, I’m American by academic predilection, which of course makes me German (B-46)

**(45)**
(*later in presentation*) And in my dissertation I focused on Germany and the UK, not just because I'm half German and half English (A-17)

**(46)**
I come from Switzerland but my parents come from Portugal. (D-01)

- *descriptions of origin: region/continent*

**(47)**
I'm a lawyer, from Asia and also educated in Asia. (D-26)

- *descriptions of origin: mobile status*

**(48)**
I'm originally from Bulgaria but I went to undergraduate school at (??) College in (??). (D-04)

**(49)**
As most of you will know by now I'm originally from Germany but I did my undergraduate degree in law and French at the University of Edinburgh. (D-12)

**(50)**
I'm originally from Israel. (D-39)

**(51)**
I'm originally from Turkey…I'm originally from (Ankara?). (B-31)

**(52)**
My name's [NAME-SURNAME] and I'm originally from south-western Germany. (D-20)

**(53)**
I'm originally from Hamburg in Germany. (A-39)

**(54)**
I'm originally from the Washington D.C. area and more recently from Northern California. (A-36)

**(55)**
I was born in England in Newcastle, but I grew up in Switzerland where I spent my formative years. (D-15)

**(56)**
My name indicate, indicates that I'm of Greek origin, but I came through the Netherlands, through Germany to the Netherlands, based in Amsterdam. (B-17)

**(57)**
I'm [NAME-SURNAME] and I am from Stockholm, Sweden, although I spent a lot of time in the U.S. (D-03)

The patterns identified suggest that, although the programme draws together scholars from many countries, national identity is not the most relevant mode of self-classification for many fellows. While 26 fellows do mention their nationality or country of origin, 7 of them immediately add their city of origin; another 15 prefer to directly invoke a local rather than national identity. Even more striking is the presence of a third group of speakers (in all, 17) who explicitly claim transnational status, either on the basis of dual nationality or dual ethnicity ("I'm half-X, half-Y" or similar construction) or on the basis of crossborders mobility (e.g. "I'm originally from..."; "I was born in England in Newcastle, but I grew up in Switzerland..."). As I will attempt to show in the following sections, these alternative modes of self-categorization are not chance occurrences. On the contrary, when examined in context they can be seen as merely the most explicit manifestations of a type of identity work being done in this setting: what, following Hougaard (2008), I will call "doing being international".

## 4.2. Positioning self as a 'transnationally mobile scholar' through narrative descriptions

In the previous section I highlighted how speakers in my data claim transnational status through self-labelling in terms of dual nationality/ethnicity or by explicitly signalling their mobility through the use of (i) the locution *I'm originally from...* and/or (ii) an adversative structure consisting in a description of origin + *but/although* + a description of subsequent mobility.

Signalling mobile status can also be achieved implicitly over a series of turns and this, in fact, is the more common strategy. The majority of the presentations open, in fact, with a short chronological narrative that showcases how the speaker's previous academic experience has been characterized by geographical and institutional mobility. Most of these narratives share certain recurrent characteristics, of which the first of those

listed below is a constant for almost all the speakers and the others, optional.[5]

**(1)** *the use of name-only reference for top-tier Western institutions, as opposed to name + geographical descriptor for lesser known or non-Western universities*. The use of name-only i.e. recognitional reference, by implying tacit knowledge about the geographical location of the former group of institutions, implicitly indexes self and audience as 'international scholars in the know' (compare example 58, in which an acronym is used for London School of Economics and name-only mention for Yale, with examples 59 and 60, in which the locations of Duquesne and Nankai universities are specified):

**(58)**
I studied at the University of Edinburgh and then I went back to Holland to do an MA degree in history and law at the University of Leyden. During that time I also spent some time at LSE and after finishing I spent a Fulbright year at Yale, where I was at the history department of the law school. (D-40)

**(59)**
I did my masters at- (*points to slide*) that's pronounced 'Duquesne University', which is in Pittsburgh. (D-03)

**(60)**
I've done my BA at Nankai University in Tianjin, a very big city in China, in the field of finance and banking. (B-15)

**(2)** *preliminary glosses which frame what is to come as a 'mobility narrative'*:

**(61)**
As many of my colleagues here, I've studied in various places: in Germany, in the U.S., in England and in France. (A-03)

**(62)**
I'm mostly Dutch and my academic background is a little peripatetic, I guess. (A-40)

**(63)**
Okay, here's my educational background. As I was getting this together I realized you know my background's very geographically diversified. (*audience laughter*) (D-36)

**(3)** *lexical choices and asides in which openness to mobility is framed in positive terms* (in the following example, *got stuck* suggests that it would have been preferable to have moved on to another institution for the next degree):

**(64)**
I'm originally from Hamburg in Germany. I studied in Venice (??) political science and economics and then I went to the LSE, where I did a Masters degree in European politics and policy and I *got stuck* in the LSE and did a PhD there at that time on civil service reform in post-communist Hungary. (A-39)

**(4)** *humour which draws on one of two geographically-based tropes (or both) for its effectiveness*: the 'itchy feet' trope (in which the speaker portrays himself[6] as always ready to move on) and the 'hometown nostalgia' trope (in which the speaker expresses nostalgia for or pride in his/her place of origin). In the second of these tropes, self-categorization in terms of local identity (see section 3.1, above) plays a prominent role. The implication conveyed by both tropes is that the speaker is a robust itinerant scholar, one who is resilient enough not to get 'lost in transit'.

*'itchy feet' trope*:

**(65)**
Okay, so my background is that I did my PhD in economics here at the EUI and I defended in 2003 and recently, after working as an assistant professor at the Institute for Advanced Studies in Vienna, in Austria, where I'm currently on leave. So I'm on leave and I might go back or I might go somewhere else, I don't know yet. (*audience laughter*) (A-13)

**(66)**
My educational background. I tried to study economics at Bocconi University, but one of the afternoons I was watching Braveheart and I decided to go study at Aberdeen University in Scotland. (A-29)

**(67)**
First, okay, my background is I'm from a small town in Ohio (*points to map on screen*), which I wanted to get the hell out of there so I did and I went to a different small town in Ohio, even smaller actually, which is (harder to find but) (*audience laughter*) And so I studied economics there, then I went to Washington D.C. for a while to work in trade policy (.) analysis (.) That was sort of boring I have to say (.) e::rm well not boring, it was interesting but I wanted to move on and did a master in economics at Pompeu Fabra and right away started to learn about some things like

social choice and some really interesting stuff about political economy, and so I got into political science. (D-29)

*'hometown nostalgia' trope*:

**(68)**
I come from the most beautiful city in the world (*audience laughter*) Strasbourg. (D-43)

**(69)**
So, I'm born in Hamburg. It's my hometown. It's my favourite town. And nevertheless I left it and spent the last seven years in France, in Paris, where I came through an Erasmus. And then I stayed there to do a master, and hopefully I will also earn my PhD there in international economics at the University of Paris 1. (C-44)

**(70)**
My research area in the past years have been business human rights law or human rights law and business [...] But before I start to talk about that, first about a thing which defines me the most, more than my academic qualifications. I'm from from Slovenia (*slide with photos of lakes at sunset appears on screen*), the most beautiful European country, and my hometown's Ljubljana (*2nd slide with photos of Ljubljana appears*), a great place for weddings. (C-29)

**(5)** *intertextual links to previous narratives by other fellows*. A number of these links emphasise similarities with other speakers by thematizing mobility, nostalgia for one's place of origin or both; in the following excerpt (example 71), the speaker is referring back to the comment made by a colleague the previous day (C-29; see example 70, above).

**(71)**
Well in Bournemouth I did my masters in philosophy and then I went to Oxford where I did my DPhil. And both places are also recommended venues for weddings. (C-37)

Not all of the scholars participating in the post-doctoral program have backgrounds that conform to this 'transnationally mobile scholar' script, but the ways this latter group of scholars present themselves are equally of interest. The following three examples illustrate how, when transnational mobility cannot be demonstrated, it is often honoured in the breach. The first two (72, 73) accomplish this through light self-denigration, the third (74), more openly by claiming that intellectual diversity can also be achieved through institutional mobility on the national level:

**(72)**
As you can hear I'm French and so about my background. I did a PhD in economics in the University of Paris Pantheon-Sorbonne. To be really honest actually I'm born in Paris so I did my school in Paris, my BA, MA, so now it's a great opportunity for me to breathe another air here. So it's really good for me, I think so. (B-39)

**(73)**
But let's start with let's say a few things about me and about my education, which was actually very parochial, because it began in Pisa, so very near, close to Florence, where I got a law degree from the University of Pisa of Pisa, then at the University of (??) and my PhD in law, curriculum in European Union constitutional law from the Scuola Superiore Sant'Anna, which is a kind of a school of advanced studies. Sure I spent I mean I also spent many months abroad I mean as a visiting fellow. (D-14)

**(74)**
I graduated at the University of Pisa, then got a PhD at Federico II University of Naples. It was important for my training and education to be in contact with these two Italian universities, not only because I faced two academic environments but also from a scientific point of view, because the historiographical approaches that are carried out in these two Italian universities are different. And so it was important to me to have the chance to face with both of them. (A-27)

Two final examples illustrate how speakers can subvert the "greater mobility = higher quality" equation by highlighting its absurdity when carried to extremes. In both cases the speakers accomplish this effect by flouting the Gricean maxim of quantity. In the first case (75), the speaker omits mention of the degrees she has obtained on the pretence that having obtained them all at the same university makes them not worth mentioning; she thus conveys irony by providing (in Gricean terms) too little information. In the second case (76), the speaker repeats *Berlin* more than would be necessary to maintain textual cohesion, thereby conferring slightly ironical overtones to his description of his educational background.[7] In both instances the audience reacts appreciatively:

**(75)**
I've done all my degrees in Tel Aviv university. So I will just skip, not present all these degrees (*laughter by both speaker and audience*) (B-45)

**(76)**
I'm pretty much Berlin-based. I was born in Berlin and did my master in Berlin, and received my PhD or will receive my PhD I hope from Humboldt University in Berlin (*chuckles from audience*). (B-18)

In the following two sections we will now move on to examine two sequential practices that involve 'altercasting', i.e. using the audience as a resource to position oneself within the virtual world of international academia.

## 4.3. Positioning self as 'intercultural' by attributing the possession of national stereotypes to the audience

A first and slightly puzzling practice consists in the activation by speakers from certain countries of sociolinguistic stereotypes regarding their own national or linguistic background. These stereotypes appear at the very beginning of the speakers' presentations and are of two types.

A first group regards difficulties the audience is presumed to encounter in understanding and remembering the speaker's name. Specifically, the audience is construed as having problems with the spelling and pronunciation of Eastern European names (examples 77-79); with names containing phonemes or phonetic realizations of phonemes not present in English, such as Italian /ʎ/ (the pronunciation of 'gli' in intervocalic position) or French /j/ (examples 80-81);[8] with Asian names in general (examples 82-83):

**(77)**
So it's pronounced György Juhász. Okay? I know it's hard. (C-42; Hungarian)

**(78)**
(*opening utterance unclear*) (*audience laughter*) by the way, the problem for me normally is that the letters get mixed up, the order is, I don't know why it's (difficult). (C-14; Hungarian)

**(79)**
My name is Blazej and last name Kozlowski, very well pronounced by the way. (B-18; Polish)

**(80)**
I'm Pierluigi Smeriglio. I:: my name is almost impossible to pronunciate (*slight audience laughter*) for the majority of you. (.) (C-16; Italian)

**(81)**
So, I'm Alain Darolle (??) not so easy to pronounce, um: French. So, I'm French. I come from Paris and well I also had the privilege to have a two-year contract so I was also a Max Weber fellow last year. (C-7)

**(82)**
Thanks for the introduction, particularly pronouncing my name. It's not easy. Actually they've offered (??) to correct our name when our name is being called in the reception of the economics department, but I haven't done that, so now we use this chance first to do it. Basically it's Fang, it's not my name there, it's the other way around (*slight audience laughter*). In China we first say family name and then forename. So it's Lu Fang, but I'm in Europe so just call me my first name Fang. It's difficult. People from America try to say [fæng], and people from Italia try to call me [fan-ge]. It's just something different but very easy, just [fäŋ].[9] But that doesn't matter. Don't hesitate to call me my name, just call what you think is more close. Okay, so it's clear I come from China. (B-15)

**(83)**
My name is Ji-Eun Han but you can call me Henry. (A-06; Korean)

The context in which the interactions take place, i.e. an EU-funded program in Western Europe in which English is the working language, may go some way towards explaining the depiction of international academia that emerges here in terms of 'centre' and 'periphery'. It begs the question, however, as to why scholars would choose to represent such a basic aspect of their personal and professional *persona* as problematic in the first place.[10] At first glance, in fact, doing so would seem to merely exacerbate the current power imbalances of contemporary academia.

I would like to suggest a possible explanation: in invoking such sociolinguistic stereotypes, speakers situate themselves as social actors who possess a sort of intercultural competence that is highly desirable, i.e. an ability to see themselves and their languages/cultures from the point of view of *other* international scholars. This interpretation is in line with work by scholars on intercultural communication from a MC perspective: in his work on conversations between Japanese and Korean speakers, for example, Zimmerman (2007) shows how the latter provide evidence of being able to see things from their interlocutor's perspective by producing descriptions that show familiarity with negative stereotypes about Korean culture. In a similar vein, the speakers in my data position themselves as interculturally savvy and 'in the know' about international academia by showing an awareness of the filter through which their audience (or at least some members of it) may be perceiving their personal information.

A second sociolinguistic stereotype emerging from the analysis is quite specific, and precisely for this reason merits attention: the notion that a French accent is easily recognizable by international scholars (in addition to the following examples, also see (8), above):

**(84)**
I'm very happy to have the opportunity to- to introduce myself and present my work in front of you. As an introduction I might say a few words about myself. I guess I don't need to mention that I'm French, because my accent speaks for itself. Some of you already know that I come from Grenoble, well I did my PhD in Grenoble, France, and in Berlin as well. (B-11)

**(85)**
Let me introduce myself. I'm French, I'll do the best to hide it. I'm a lawyer, I should try to hide this as well probably. And as many of my colleagues here, I've studied in various places: in Germany, in the U.S., in England, and in France. (A-03)

**(86)**
Okay. So first of all I will say that today is my first big speech in English so I will be a little nervous. (*audience laughter*). But after hearing presented many speeches last Friday in Fiesole, I think that you will excuse my French accent. (C-24)

Why French, but none of the other native and non-native accents of English to be heard in this highly linguistically diverse setting? And again, why should a number of French scholars – who, after all, come from a country in the heart of Western Europe – choose to evoke a stereotype that in a certain sense sets them apart from the rest of their peers?

Studies examining the links between national language policies and language use by European academics indicate that publishing and presenting in the national language remain central to career advancement in France, with the use of English less highly valued than in most other European countries (Wright 2006; Marimon et al. 2009; Anderson 2013).[11] French has traditionally had, moreover, an important role as a lingua franca in the European context and continues to be promoted by the French government for use in crossborders communication between French scholars and Francophone scholars from other regions. In invoking this stereotype, these French scholars may, whether consciously or not, be indirectly reminding their audience of the existence of this alternative arena of academic interaction. In this interpretation, underlining their French accent when speaking in English may be one way by which they

can signal the value of the French academic tradition – or, at any rate, of linguistic plurality – as an integral part of the European academic scene.

### 4.4. Positioning self by altercasting the audience as a member of a geopolitical relational pair

As we have seen above, evoking a sociolinguistic stereotype is one way in which speakers position themselves by altercasting the audience: in doing so, they construe the audience as a microcosm of international academia and demonstrate ‘internationality’ by showing an awareness of ‘the lay of the land’. The audience can also be altercasted by portraying it as a member of a relational pair. Sequences of this sort generally occur later in the presentations when speakers are describing their research. Here, I will present four examples that illustrate how speakers in my sample use this strategy to position themselves as *specific types* of international scholars.

The first two excerpts illustrate a procedure used by some scholars who received their doctoral training in North America. It involves altercasting the audience as generically ‘European’. In example (87), a researcher attributes to the audience expertise in work on aging carried out within the European context. By doing so she makes relevant the relational pair ‘European/North American’ and positions herself in the complementary role of ‘North American scholar’. In example (88), instead, the presenter attributes insider knowledge about the European job market to the audience; this makes it possible for him to frame himself contrastively as an expert on the American one:[12]

**(87)**
I would appreciate any feedback that you have on aging research in Europe (A-01)

**(88)**
Now, since I’ve been out for a while, I probably had some experiences that you, some of you have had, will have had, and some of you will have in the next couple of years. Been on the job market twice. I know it differs from Europe in the United States. But I can, you know, I can give you some points or some- how to have awkward chit-chat with faculty, you know, how to talk to a dean, maybe even a president. Happy to share those experiences with you. I’ve been on search committees too, which makes you wonder how anybody ever gets a job. So if you wanna know how that process works I can talk about it with you. And I wanna know more about the European market, which is something I really don’t know about. (B-03)

It should be stressed that in using the audience as a backdrop against which to frame their own identity, presenters are not simply reflecting some sort of pre-constituted reality. It is possible, in fact, to construe the audience in various ways. This can be clearly seen in the following two examples, both involving speakers from Eastern European countries which are not members of the European Union – respectively, the Ukraine and Russia. In example (89) the speaker altercasts the audience as 'scholars from the European Union', thus activating the relational pair 'European Union vs. non-European Union scholars'; in example (90), instead, the presenter frames the audience as 'scholars from Western Europe' (example 90), thus activating the relational pair 'Western European vs. Eastern European scholars':

**(89)**
Probably for people from the European Union it is much easier but for us Ukrainians and for us, for people from other countries it's not, it's really a challenging exercise (A-20)

**(90)**
So there were lots of works of Russian scholars which are little known here in Western Europe and which were used in comparative research with, well, classical or less known works by West Europeans. ..... Well the, probably it may be strange for you but the peculiarity of Russian legal tradition, Russian legal thinking, as well as political history, is informed by a very strong concern for sovereign immunity, which even now partly explains the tradition of seeing European Community or European Union as an international organisation. I think that cultural and linguistic value of this project should be interesting as well probably for you. (A-40)

What is the utility to the speaker of altercasting the audience in geopolitical terms? In scrutinizing these and similar examples, what is striking is how academia is depicted in terms of – to borrow a catchword currently much in vogue in EU circles – a "knowledge economy": the audience and speaker are framed as members of complementary categories that have something to learn from each other. Not surprisingly, given the setting, in my data speakers tend to categorize the audience as 'European', but how broadly this category is construed and, above all, on what 'trade axis' it is seen as operating – North Atlantic, EU/non-EU, Western Europe/Eastern Europe – is open to negotiation. By activating a categorization device (Sacks 1972b) that best suits their individual situation and research profile, presenters can enhance their status as bonafide and useful members of international academia.

## 5. Concluding remarks

In the analysis presented in this study, we have observed how, in peer-to-peer interaction with colleagues from different national and linguistic backgrounds, the scholars in my data exhibit *a particular understanding of international academia*. It is a transnational space characterized by well known landmarks (the institutions of higher education referred to through recognitional reference), a centre (or centres, as the French scholars seem to discretely remind us) and peripheries, a space to be navigated by scholars who are mobile by choice or as a result of the globalizing dynamics of contemporary academia. At the same time, it is also a 'knowledge marketplace', one in which groups of researchers – variously construed in geographical and geopolitical terms – exchange the results of their research, constructing networks and laying the foundations for future collaboration. This understanding of international academia is, moreover, *context-sensitive*, reflecting both the geographical/geopolitical and institutional characteristics of the setting in which the presentations took place (an EU-funded training initiative located in Western Europe).

A second group of observations regards recurrent patterns in the practices of self- and other-categorization identified and, more broadly, in the "narratives of 'me'" (Mauranen 2013) in which the presenters engage. We have seen how these *speakers situate themselves not only at, but also below and above/beyond, the level of the nation-state*. References to nationality and national origin are used to thematize both hybridity ("I am part-X, part-Y") and mobility ("I am *originally* from X"); references to cities, towns and other sub-national geographical units serve both to construct trajectories in time and place and to colour them with the emotional overtones familiar to the migrant experience (nostalgia, desire for change). The picture of face-to-face communication among academic peers that emerges from the analysis thus aligns closely with the description of ELF communication quoted in the opening section of this chapter, i.e. a mode of communication whose distinguishing features are "non-locality, non-permanence, speaker mobility, and multilingualism" (Mauranen 2012, 23).

The scholars examined in the present study are successfully handling the challenges of globalizing academia, as evidenced by their participation in a highly-selective post-doctoral programme; for this reason, scrutiny of their discourse practices can offer useful insights for EAP practitioners. With the changing structure of academic careers, today's graduate students, post-docs and junior faculty – particularly those from non-Anglophone countries – increasingly find themselves working and

interacting in ELF contexts as a result of either short-term or long-term mobility. Communicating successfully face-to-face in international settings is important, however, even for scholars already established in national university systems: participating effectively in conferences and other occasions in which research is presented orally is essential for constructing research networks and for accessing the publishing opportunities needed for career advancement. In short, although some of the ways of "doing being international" documented in this study may be particularly pertinent for mobile, early-career scholars, positioning oneself appropriately on the international scene is increasingly necessary for academics at whatever stage of their careers.

What the present study shows is that EAP practitioners and the scholars they work with can profit considerably from viewing academic presentations as not just an informative but also a relational genre. As we have seen, these two aspects of communication are intimately intertwined, not only in the opening sections of presentations but also in framing one's research and highlighting its pertinence to listeners from other national contexts and linguistic backgrounds. EAP learners and ELF scholars should be alerted both to *where* in academic presentations such 'identity work' is typically done and to *what linguistic and pragmatic resources* are generally drawn on to carry it out.[13] As we have seen, such resources include not only typical ways of discretely showing that one is 'in the know' (about the notoriety of institutions of higher learning; about the lens through which scholars from other national settings and cultural backgrounds may be experiencing one's speech and presentation style) but also, and more importantly, ways of activating categorization devices – in the form of relevant relational pairs – that frame one's position within international academia to best advantage.[14] Membership categorization, even in the rather simplified form in which it has been employed in the present analysis, is an analytic tool that can prove useful to both EAP teachers and the academics with whom they work, thanks to its capacity to foreground the ways in which academic presentations feed into and support a range of social activity types in the lives of scholars.

## Notes

[1] In this emerging scenario of globalized academia, English as an academic lingua franca obviously has a central role to play – one that, given the geopolitical forces involved, is by no means neutral. As is under the eyes of all those interested in the sociolinguistics of contemporary English, a key (and often contentious) development in recent years has been the emergence of the English as a Lingua Franca (ELF) approach to academic English, which, in proposing a 'users as

owners' model of language (Haberland 2011), is sometimes seen as in contrast with the EAP (English for Academic Purposes) endeavour, sometimes accused of adopting an 'assimilationist' approach. For exploration of these issues, see Anderson (2010) and Mauranen (2012).

[2] Relational pairs can be seen as the simplest type of 'Membership Categorization Devices' (MCD) – basically, structured collections of categories. For more extensive descriptions of MC, see Schegloff 2007; Antaki and Widdecombe 1998. It should be noted that MC analysis has developed in recent years in several different directions; the distinctions between these various approaches are not considered here.

[3] I would like to thank Letizia Cirillo for her assistance in the data analysis and for many fruitful discussions about the data. A co-authored paper focusing on how membershipping practices contribute to establishing and maintaining disciplinary communities and to building bridges between them is currently in preparation.

[4] Each of the four cohorts is referred to by a letter (A, B, C, D); the number refers to the order in which the presentation took place in the year in question.

[5] For an illustration of how a constellation of recurrent characteristics contributes to "doing being international" in another context, see Hougaard (2008, 323-24) on narrative monologues in crossborders business telephone calls.

[6] Interestingly, the 'itchy feet' trope is only used by male speakers.

[7] Note the use of name + geographical descriptor to refer to Humboldt University, an institution which, as highlighted above, is as a general rule referred to simply by name-only.

[8] To protect the speakers' anonymity, proper names have been substituted with others indexing the same national and linguistic background; linguistically salient elements, such as the presence of given phonemes, have been preserved.

[9] The phonetic transcriptions are approximate and intended simply to indicate the contrasts highlighted by the speaker.

[10] Stereotypes regarding difficulties in accessibility are the most common, but speakers can also frame their name as recognizable: "My name indicates that I'm of Greek origin". (B-17); "Well, I'm Ukrainian, as you can- some of you can probably tell from my name."(B-28). In the light of a general tendency to frame Eastern European names as difficult, it is interesting to note the presence in the latter example of a false start, after which the speaker restricts the scope of her claim.

[11] Economics and the hard sciences constitute a partial exception. On the tension experienced by French academics due to the contrast between legislation mandating the use of French in higher education and use of English in international publishing, see Wright (2006); on the impact of this situation on language attitudes and publishing behaviour by early-career French scholars, see Anderson (2013).

[12] Situating himself as 'American' by altercasting the audience as 'European' is a recurrent feature of B-03's presentation. For another example, see example (14) quoted in section 3.1, from earlier in the same presentation.

[13] The present contribution focuses on self- and other-categorization in geopolitical terms; the other key dimension along which categorization occurs in academic

discourse is disciplinary membership. In a teaching perspective it will be essential to take categorization devices of both types into consideration.

[14] Curry and Lillis (2010) have highlighted how strategically positioning one's research in terms that makes it relevant to the international community is important in negotiating access to publishing opportunities in international journals.

## References

Aalbers, Manuel B., and Ugo Rossi. 2007. A coming community: Young geographers coping with multi-tier spaces of academic publishing across Europe. *Social and Cultural Geography* 8(2): 283-302.

Anderson, Laurie. 2010. Standards of acceptability in English as an academic lingua franca: Evidence from a corpus of peer-reviewed working papers by international scholars. In *Discourse, communities, and global Englishes*, ed. Roberto Cagliero and Jennifer Jenkins, 115-144. Bern: Peter Lang.

—. 2013. Publishing strategies of young, highly mobile academics: The question of language in the European context. *Language Policy* 12(3): 273-288.

Antaki, Charles, and Sue Widdecombe. eds. 1998. *Identities in talk*. Thousand Oaks, CA/London: Sage.

Benwell, Bethan, and Elizabeth Stokoe. 2006. *Discourse and identity*. Edinburgh: Edinburgh University Press.

Björkman, Beyza. ed. 2011. *The pragmatics of English as a lingua franca in the international university.* Special issue of *Journal of Pragmatics* 43.

Charle, Christophe, Jurgen Schriewer, and Peter Wagner. eds. 2004. *Transnational intellectual networks: Forms of academic knowledge and the search for cultural identities*. Frankfurt/New York: Campus Verlag.

Day, Dennis. 1994. Tang's dilemma and other problems: Ethnification processes at some multicultural workplaces. *Pragmatics* 4(3): 315-336.

Haberland, Hartmut. 2011. Ownership and maintenance of a language in transnational use: Should we leave our lingua franca alone? *Journal of Pragmatics* 43: 937-949.

Hester, Stephen, and William Housley. eds. 2002. *Language, interaction and national identity: Studies in the social organisation of national identity in talk-in-interaction*. Aldershot: Ashgate.

Hougaard, Gitte R. 2008. Membership categorization in international business phonecalls: The importance of 'being international'. *Journal of Pragmatics* 40: 307-332.

Jenkins, Jennifer. 2007. *English as a lingua franca: Attitude and identity*. Oxford: Oxford University Press.

—. 2014. *English as a lingua franca in the international university: The politics of academic English language policy.* London/New York: Routledge.

Kim, Terri. 2009. Shifting patterns of transnational academic mobility: A comparative and historical approach. *Comparative Education* 45(3): 387-403.

Leydesdorff, Loet, and Caroline S. Wagner. 2008. International collaboration in science and the formation of a core group. *Journal of Informetrics* 2: 317-325.

Lillis, Theresa, and Mary Jane Curry. 2010. *Academic writing in a global context: The politics and practices of publishing in English.* London: Routledge.

Marginson, Simon. 2007. Have global academic flows created a global labour market? In *World yearbook of education 2008: Geographies of knowledge, geometries of power: Framing the future of higher education*, ed. Debbie Epstein, Rebecca Boden, Rosemary Deem, Fazal Rizvi and Susan Wright, 305-318. New York: Routledge.

Marimon, Ramon, Matthieu Lietaert, and Michele Grigolo. 2009. Towards the ‘fifth freedom’: Increasing the mobility of researchers in the European Union. *Higher Education in Europe* 34(1): 25-37.

Mauranen, Anna. 2012. *Exploring ELF: Academic English shaped by non-native speakers.* Cambridge: Cambridge University Press.

—. 2013. “But then when I started to think...”: Narrative elements in conference presentations. In *Narratives in academic and professional genres*, ed. Maurizio Gotti and Carmen Sancho Guinda, 45-65. Bern: Peter Lang.

Nishizaka, Aug. 1999. Doing interpreting within interaction: The interactive accomplishment of a “henna gaijin” or “strange foreigner”. *Human Studies* 22: 235-251.

Sacks, Harvey. 1972a. An initial investigation of the usability of conversational data for doing sociology. In *Studies in social interaction*, ed. David N. Sudnow, 31-74. New York: Free Press.

—. 1972b. On the analyzability of stories by children. In *Directions in sociolinguistics: The ethnography of communication*, ed. John Gumperz and Dell Hymes, 325-345. New York: Holt, Rinehart and Winston.

Schegloff, Emanuel A. 2007. A tutorial on membership categorization. *Journal of Pragmatics* 39: 462-482.

Seidlhofer, Barbara. 2011. *Understanding English as a lingua franca*. Oxford: Oxford University Press.

Van De Mieroop, Dorien. 2008. Co-constructing identities in speeches: How the construction of an 'other' identity is defining for the 'self' identity and vice versa. *Pragmatics* 18(3): 491-509.

Wagner, Caroline S., and Loet Leydesdorff. 2005. Network structure, self-organization, and the growth of international collaboration in science. *Research Policy* 34: 1608-1618.

Weinstein, Eugene A., and Paul Deutschberger. 1963. Some dimensions of altercasting. *Sociometry: Journal of Interpersonal Relations* 26: 454-466.

Wright, Sue. 2006. French as a lingua franca. *Annual Review of Applied Linguistics* 26: 35-60.

Zimmerman, Erica. 2007. Constructing Korean and Japanese interculturality in talk: Ethnic membership categorization among users of Japanese. *Pragmatics* 17(1): 71-94.

# CHAPTER NINE

# INSTITUTIONAL ACADEMIC ENGLISH AND ITS PHRASEOLOGY: NATIVE AND LINGUA FRANCA PERSPECTIVES

ADRIANO FERRARESI
AND SILVIA BERNARDINI
UNIVERSITÀ DI BOLOGNA, ITALY

## 1. Introduction: Why institutional academic English

Within the field of English for Academic Purposes, substantial work has been devoted to academic research genres used for knowledge sharing and central in terms of scientific achievement, e.g. Ph.D dissertations and defences, research articles and talks, as well as subgenres such as research article abstracts and introductions (see Swales 2004 for an overview). In recent years we have also witnessed a surge of interest in genres relevant to other aspects of academic life, e.g. book reviews (Römer 2010), grant proposals (Connor and Upton 2004), thesis acknowledgements, doctoral prize applications and bio statements (Hyland 2011).

Such genres situate themselves midway between the strictly disciplinary genres traditionally focused upon in discourse and genre studies (e.g. the research article), and the genres used for everyday institutional academic communication – especially between institutions and their students, i.e. syllabi, course packs, welcome messages, mission statements, announcements and so forth. Probably due to their subservient housekeeping function, these institutional academic genres have so far been largely neglected as objects of study (with some exceptions that will be discussed in Section 2). Yet this state of affairs is bound to change, as universities worldwide place more and more importance on strategies for attracting prospective students and effectively managing relations with current students and alumni.

This is especially true of Europe at the moment, where academic institutions are under increasing pressure to market themselves beyond national borders. Efforts at internationalization are inherent in the *Bologna Process*,[1] which requires universities to recruit international students and attract exchange staff and students through mobility programmes. Several studies investigating higher education policy have shown that for this internationalization process to be successful, availability of academic modules and/or entire degree courses in English is essential (e.g. Altbach and Knight 2007), and that one of the most effective means to reach a vast international audience is to publish (quality) contents in English on institutional websites, which are a primary source of information for up to 84% of prospective students (cf. Saichaie 2011, Chapter 1, and references therein).

Considerable variability, however, is observed when taking into account the degree to which academic institutions of different European countries respond to this demand for (web-based) English contents. In a large-scale study on Internet multilingualism, Callahan and Herring (2012) find that the presence of English as a secondary language is most widespread on the websites of West-European (and especially Scandinavian) universities, followed by universities from the post-Soviet bloc. On the other hand, Romance-language countries like France, Italy and Spain tend to lag behind.

Against this background, interventions aimed at supporting multilingualism in the institutional/administrative domain are therefore in order. On the practical/applied side, these may include the implementation of tools for assisting non-native writers in producing appropriate texts in this specialised domain (cf. Depraetere et al. 2011); on the descriptive side, studies are required which shed light on the different communicative strategies adopted by universities based in countries where English is used as a native language or as a lingua franca.

The present chapter intends to contribute to this second line of research by pursuing two inter-related aims: first, it introduces acWaC-EU (an acronym for "*ac*ademic *W*eb-*a*s-*C*orpus in *Eu*rope"), a 90-million word corpus of institutional academic texts in English, collected using semi-automatic procedures from the websites of European universities; second, it aims to provide a preliminary characterization of the native and lingua franca varieties represented in acWaC-EU, with respect to their phraseology.

The remainder of the chapter is organised as follows. Section 2 presents the double-sided background to this work: it first discusses studies focusing on institutional academic language, and then briefly

reports on previous work investigating the features of English when it is used as a lingua franca, with particular reference to its use in academic settings. Section 3 moves on to introduce the pipeline that was followed to build acWaC-EU and provides basic information on the corpus. Section 4 then presents a case study replicating the methodology of Durrant and Schmitt (2009) to compare the phraseology extracted from native and lingua franca texts in acWaC-EU, aimed at assessing the extent to which the latter display (dis)similar patterns compared to the former in terms of use of infrequent word combinations and "strong", salient collocations. Section 5 concludes by summing up and briefly suggesting possible ways forward.

## 2. Previous studies

Landmark works focusing on institutional academic genres have been produced mainly within applied (corpus) linguistics and critical discourse analysis. The former are motivated by the observation that institutional texts constitute a substantial share of the (non-research) texts that faculty members are expected to produce as part of their commitments (Hyon and Chen 2004), as well as being required readings for students who need to "navigate the maze of university requirements and services" (Biber 2006, 26). The latter spring from concerns with the increasing tendency for universities to adopt business models and transform education into a saleable good, which are hypothesized to be reflected in their discursive practices.

Within the corpus linguistics approach, Biber (2006) provides a full-fledged account of the TOEFL 2000 Spoken and Written Academic Language corpus (T2K-SWAL), which includes both academic and institutional genres (e.g. handbooks, catalogues, programme web pages, course syllabi). These are revealed as complex hybrid texts in which different functions and styles coexist (e.g. informative and directive, personal and impersonal), to the extent that "the linguistic style found in many university catalogues and program brochures is often more reminiscent of highly technical academic prose than textbooks written for novices in an academic discipline" (Biber 2006, 189). The relevance of institutional genres for applied linguistics purposes is also endorsed by the builders of the MICASE corpus, constructed at the University of Michigan, which includes samples of spoken academic registers not limited to lectures and seminars but also including more informal, everyday events such as service encounters and campus tours (Simpson-Vlach and Leicher 2006). Recent work has also begun to explore the

specific features of different genres within the domain, e.g. course syllabi (Afros and Schryer 2009; Gesuato 2011) and the "About us" pages of university websites (Caiazzo 2011).

Moving on to the critical discourse analysis perspective, work on the discursive practices of tertiary education institutions dates back to the seminal paper by Fairclough (1993). Here, it was suggested that universities are "in the process of being transformed through the increasing salience within higher education of promotion as a communicative function", which in turn raises doubts as to "what is happening to [...] authority relations between academics and students, academic institutions and the public" (Fairclough 1993, 143). Surveying more recent trends in academic communication, Swales (2004, 9) argues that the "marketization" of university discourse has also been accompanied "by a shift in curricular perspective to the needs of the students (now seen as "customers") as opposed to the scholarly expectations of a discipline or the traditional offerings of a department". Evidence that this process is increasingly pervasive has made the object of a number of papers – some raising trenchant criticisms (e.g. Webster 2003) – within critical discourse analysis. Mautner (2005, 38), for instance, shows how universities borrow commercial models, using persuasive style and "[l]exical imports from the business domain", a finding confirmed by Morrish and Sauntson (2013, 78), who argue that institutions "have adopted the language of business and industry, managerialism and neoliberalism".

One shared aspect of the corpus linguistics and critical discourse studies reviewed so far is the adoption of a US and/or UK-centred perspective, which fails to account for the discursive practices of universities based in non-Anglophone countries. Viewing academia as "one of those influential domains that have widely adopted English as their common language, and […] where international communication characterizes the domain across the board" (Mauranen 2010, 21), explorations of non-native English varieties in international academic settings have been produced within studies of English as a lingua franca (ELF for short; see e.g. Jenkins 2011 for an overview). While focusing mainly on oral rather than written communication, these studies have brought to the fore the importance of isolating the features that characterize effective communication in ELF, and set it apart from its native counterpart. These include, e.g. a strong reliance on literal expressions rather than figurative, formulaic language (Kecskes 2007), and a relative over-production (and tolerance) of non-standard forms, both at the grammatical and lexical levels (Mauranen 2010; Jenkins 2011).

To the best of our knowledge, the only study that has set out to compare ELF and native production in the institutional academic domain is Bernardini et al. (2010).[2] The authors describe a corpus of institutional academic texts collected from the websites of British/Irish and Italian universities using a semi-automatic procedure that consisted in manually selecting relevant URLs and then using these as "seeds" for retrieving and downloading pages through the BootCaT toolkit (Baroni and Bernardini 2004). The authors compare the native and ELF sub-corpora in terms of genres and topics covered, phraseological patterns and stance expressions, and find that ELF texts are focused on spelling out instructions and requirements while native texts promote themselves as service providers through a personal style. The corpus and the case study presented in Sections 3 and 4 build on and extend this work.

## 3. Corpus construction

The acWaC-EU corpus of institutional academic texts can be considered as an enhanced (and enlarged) "2.0" version of the acWaC corpus presented in Bernardini et al. (2010). Unlike the previous corpus, acWaC-EU includes web pages of universities from *all* European countries, and its construction procedure excludes the manual browsing of web pages. Other strengths of the pipeline that was developed to build acWaC-EU include the following:

- it is designed to maximize chances that only relevant texts are included in the corpus, leaving out texts belonging to research – rather than institutional academic – genres;
- it addresses the non-trivial problem of identifying English contents in multilingual websites;
- it makes it possible to replicate corpus construction, e.g. for monitoring/diachronic purposes;
- it can be easily adapted so as to include a) the websites of universities based in countries outside Europe; b) a higher number of web pages from each university website.

Of course, an automatic pipeline like the one presented here, based on the so-called "web-as-corpus" approach (Baroni et al. 2009), is not devoid of problems, but it was deemed to achieve a favourable trade-off between scope (number of pages retrieved for each university, number of universities per European country, etc.) and quality (variety of text types

included, "cleanliness" of the texts, etc.), while ensuring replicability, e.g. for monitoring purposes or for future extensions.

The corpus construction procedure consists of three main phases: a) "seed URL" retrieval and harvesting of pages, b) post-hoc cleaning and c) tagging and indexing. Each of these is illustrated in turn.

In the first phase, a list of the homepage URLs of all European universities is obtained from the *Webometrics* website,[3] which publishes a yearly ranking of universities based on the quality of their research and teaching activities (as reflected on their websites). A Perl web crawler then visits each homepage and downloads it.

For universities based in countries where English is a native/official language (native universities for short), the URLs obtained from *Webometrics* are used to seed the second crawl (below). For all other European universities (ELF universities), the crawler first analyses the HTML code of the page and looks for a link to the English-language homepage (if any is present). This is done by means of simple Regular Expressions looking for the pattern `(english|eng|en)` (both lower- and upper-case) in the `href`, `class` and `title` attributes of `<a>` tags, and in anchor text. Links like, e.g. http://www.unimi.it/ENG/ are thus identified. Admittedly, the method is rather naïve: it also selects (supposed) English pages not actually in English, and ignores English pages with a different URL syntax (e.g.http://www.international.unina.it/). Yet it constitutes a "principled" technique to identify English content in non-English websites (a similar procedure is followed by Callahan and Herring 2012; cf. Section 1), and random manual inspection reveals that it is effective in a substantial number of cases: 2,622 links are found in this way, out of a total of 5,505 university websites.

The URLs found using this method and the native English homepages obtained from *Webometrics* are used to seed the second crawl. In this further step, the pages linked from the homes are fetched, with two levels of recursion, i.e. we download pages if they are at most two links away from the seed URLs. This maximizes the chances that only relevant texts are collected from ELF university websites, since, as one moves away from the English homepage, contents in English dwindle and the probability that the downloaded texts do not belong to the target domain (e.g. research articles, conference calls for papers, etc.) increases (Ferraresi and Bernardini 2013).

In the second phase, the crawled web pages are cleaned. This is a necessary step in web-as-corpus projects, especially when customized crawls of the web are performed (i.e. when the procedure does not rely on search engines like Google to retrieve web pages) and it is highly likely

that unwanted pages have been downloaded, e.g. duplicate texts, texts that are not in the target language, etc. The tools developed for the web-derived, general-purpose WaCky corpora (Baroni et al. 2009) were used, i.e. a language identifier, a text de-duplication algorithm, and a "boilperplate-stripping" heuristic, which extracts the text portion of web pages discarding images and formatting,[4] as well as navigation bars, headers, footers and the like, i.e. all those text fragments which tend to be repeated verbatim across all pages, and can thus distort corpus statistics. Ferraresi and Bernardini (2013) report high levels of precision for the corpus building procedure: the proportion of texts matching the target population (published by a university, not containing machine-generated text etc.) is estimated to be above 90%.

After cleaning, the texts are part-of-speech tagged and lemmatized using the TreeTagger[5] and indexed for consultation with the Corpus Workbench.[6] During this phase, a rich layer of contextual metadata are recorded with each text (corresponding to a single web page), including:

- URL of the web page;
- level in the site structure in which it was found (from 0 to 2, where 0 indicates the English homepage);
- name of the university it belongs to;
- university rank according to the *Webometrics* ranking;
- country;
- status of English in the country (native or ELF);
- language family of the official language spoken in the country (e.g. Romance, Germanic, Slavic, etc.).

These metadata can be exploited to build subcorpora, e.g. to compare native vs. ELF texts, ELF texts from countries where the official language is Romance vs. Germanic and so forth.

Table 1 presents summary statistics on the acWaC-EU corpus, split according to whether the texts belong to the native (NAT) or to the ELF subcorpus. Further information on the corpus and updates on its development are available from the project website: http://mrscoulter.sslmit.unibo.it/acwac/.

| | NAT | ELF |
|---|---|---|
| Number of tokens | 46,172,429 | 41,696,310 |
| Number of texts | 68,011 | 73,296 |
| Number of universities | 341 | 2,159 |
| Number of countries | 4 | 46 |

**Table 1. Summary statistics for the acWaC-EU corpus.**

# 4. Case study: Phraseology in native and lingua franca institutional academic English

## 4.1. Introduction

This case study investigates the use of phraseology in (a subset of) the native and ELF subcorpora of the acWaC-EU corpus. This seems to be a particularly promising area of investigation in the exploration of similarities and differences between the two varieties. Starting from the seminal article by Pawley and Sider (1983), the use of phraseological items – also variously called "collocations", "lexicalized phrases", "stock phrases", "formulaic sequences", the list could go on – has been among the most widely researched topics in studies of learner language (see e.g. Nesselhauf 2005 and Meunier and Granger 2008). The widely shared view is that learning and (appropriately) using "prefabricated sequences of words" – adapting Wray's (2002, 9) definition – constitute key points of difficulty for non-native speakers, even at advanced levels of proficiency (Nesselhauf 2005). Simplifying somewhat, it is hypothesized that non-native speakers use phraseology to a lesser extent than natives, and that the former tend to produce more deviant/non-standard/unidiomatic sequences, i.e. what Pawley and Sider (1983, 191) call "non-nativelike or highly marked usages".

Durrant and Schmitt (2009) describe a promising methodology to investigate the (dis)similarity between native and non-native texts in terms of phraseological patterns. Comparing the written production of learners of English vs. native English speakers,[7] the two authors focus on collocations "as they have been defined by corpus linguists of the 'neo-Firthian' school", i.e. as "words which appear together in the language more often than their individual frequencies would predict" (2009, 159). Their method is appealing for two reasons. First, it sets objective parameters for the identification of word combinations in corpora, based on their values of frequency of co-occurrence and of two lexical association measures (*t*-score and Mutual Information). Second, it relies on frequency data gathered from a large reference corpus (in their case the

British National Corpus, or BNC),[8] which makes it possible to tell apart word combinations which "have common usage in English" in general (2009, 167) from those that are idiosyncratic to the corpus under examination, a problem faced by several studies of learner language (e.g. Nesselhauf 2005; cf. Durrant and Schmitt 2009, 160). It is thus possible to perform a direct comparison between the "degree of formulaicity" of native and ELF texts, while at the same time limiting researcher bias in selecting what word combinations constitute a collocation.

Section 4.2 provides a detailed account of the method that was followed in the present study, highlighting the changes made with respect to Durrant and Schmitt's (2009; henceforth DS) method. Section 4.3 then presents the results and discusses their implications for the description of native and lingua franca English in the institutional academic domain.

## 4.2. Method

As is the case with virtually all comparisons across different (sub)corpora/language varieties/etc., comparing phraseology across texts produced by native and lingua franca speakers requires that several variables such as text genre and length are controlled, so as to limit their influence. The first step consisted therefore in identifying appropriate comparable texts in the NAT and ELF subcorpora of acWaC-EU. Given the narrower scope of this case study compared to DS (cf. note 7), the focus was on a single text type, i.e. homepages.

One might argue that homepages in the NAT and ELF subcorpora are targeted at a different audience, i.e. at a rather homogeneous public of national students and staff vs. a highly diversified international audience, thus raising doubts as to their actual comparability in terms of communicative function. It is reasonable to expect that ELF homepages are more informative than NAT ones, since their audience is not necessarily familiar with the education system of the country, to mention but one aspect. As DS note, however, "[i]dentifying native texts that are equivalent in type to non-native writing is […] highly problematic" (DS, 162). Homepages at least have the advantage of being clearly identifiable using "text external" criteria (cf. Sinclair 2004), i.e. their URL.

Following DS, 24 texts were selected among the NAT and ELF homepages, for a total of 48 texts.[9] Preliminary inspections revealed that NAT homepages were fewer than ELF ones and that they tended to be shorter. Since the method proved less robust for short texts (DS, 162), it was decided to take the 24 *longest* NAT texts, and to randomly select ELF texts of comparable length. Exploiting acWaC-EU metadata, the further

constraint was imposed that ELF texts had to be from Romance language countries (cf. Section 3). This was done to control for the influence of different first language families on the phraseology produced in a second language (Wolter and Gyllstad 2011). Table 2 reports summary statistics about the selected texts.

| | NAT | ELF |
|---|---|---|
| Number of texts | 24 | 24 |
| Number of universities | 24 | 24 |
| Total words | 14,372 | 13,911 |
| Mean words/text (+ Standard Deviation) | 599 (SD: 311.13) | 580 (SD: 310.37) |
| Countries | UK (14)<br>Ireland (10) | France (11)<br>Italy (6)<br>Spain (3)<br>Portugal (2)<br>Romania (1)<br>French Switzerland (1) |

**Table 2. Texts selected for the case study: summary statistics.**

The extraction and analysis of phraseological items was limited to a single syntactic pattern, i.e. "directly adjacent premodifier-noun word pairs" (DS, 162). Instances of adjective-noun and noun-noun combinations were extracted, relying on part-of-speech information, and the same filtering steps as in the original study were applied, i.e. combinations were excluded if they contained proper nouns (as identified by the tagger; this also excluded most acronyms), semi-determiners (i.e. *same*, *other*, *former*,*latter*, *last*, *next*, *certain*, *such*) and numbers; word pairs containing ordinals were instead retained, to keep track of phrases like, e.g. "*first/second/third year*", which are arguably central to the (institutional) academic domain. All word combinations were then lowercased, and the frequencies of the upper- and lowercase variants collapsed, for a total of 910 combinations in the NAT subcorpus and 1,170 in the ELF one.

Information on the frequency of the extracted word combinations was then gathered from a general reference corpus. This is a crucial step, which makes it possible to assess the extent to which phrases are common and/or salient in contemporary English in general, rather than in the corpus under examination. In practical terms, all word combinations – even those with frequency equal to 1 in the NAT/ELF texts – were considered as potential collocation candidates, and their frequency was checked in a subset of

ukWaC, the English WaCky corpus (Baroni et al. 2009; cf. Section 3). This corpus was constructed more than ten years after the BNC, and contains texts more comparable to the ones in acWaC-EU: for instance, the phrase "*international students*" only occurs 3 times in the BNC, and hence would have been considered by DS' parameters as a low-frequency, non-salient combination. Since we wanted to apply the same parameters for collocation identification as in DS, a random sample of 180 million words taken from ukWaC was used.[10]

As in DS, information on raw frequency of co-occurrence in the reference corpus (henceforth FQ) was used to calculate association scores according to two lexical association measures, i.e. *t*-score (henceforth *t*) and Mutual Information (henceforth MI). *t* values were calculated using the UCS toolkit,[11] and MI values using an ad-hoc Perl script implementing the formula by Church and Hanks (1990). The two association measures give prominence to different types of salient word combinations: *t* highlights "very frequent collocations" (DS, 167) like, e.g. "*more information*", "*higher education*" and "*third year*", while MI selects "word pairs which may be less common, but whose component words are not often found apart" (DS, 167), e.g. "*land-based industries*", "*entrance examinations*" and "*senior lecturer*". Association scores were computed only for pairs with a frequency of at least five occurrences in ukWaC.

In the last step, again following DS, for *each* NAT and ELF text the following figures were calculated:

- percentage of low frequency word combinations (i.e. those with FQ < 5 in ukWaC);
- percentage of highly salient combinations according to *t*-score values ($t \geq 10$ in ukWaC);
- percentage of highly salient combinations according to MI values (MI ≥ 7 in ukWaC).

Percentages were calculated as the number of word combinations occurring in a single text and satisfying one of the above conditions out of the total number of word combinations in that text.[12] Compared to previous analyses of native vs. non-native writing, this method has the advantage that it does not "disguis[e] differences between individual texts" and is therefore less likely to "produce misleading results" (DS, 168). In an analysis, e.g. of low-frequency combinations (cf. Section 4.2) using a "whole-corpus approach" one would calculate percentages based on the total number of word combinations in the NAT vs. ELF subcorpora (cf. DS, 168–169). Percentages were instead calculated for each of the 48

texts, and then averaged across the 24 texts in each subcorpus. This makes it possible to

> use inferential statistics to find whether texts of one type contain a significantly higher percentage of infrequent collocations than those of another. Significant scores on these tests will indicate relative homogeneity within groups and meaningful differences between them. (DS, 179)

Section 4.3 presents and discusses the results of the procedure.

## 4.3. Results and discussion

The first part of the analysis focused on the extent to which NAT and ELF texts rely on word combinations which are infrequent in English, defined as the pairs with FQ < 5 in ukWaC. Figure 1 displays the distribution of the results, i.e. the percentages of infrequent pairs *in each text*: the thick black lines in the plot represent median values, the grey boxes the 50% of data points around the medians and the top and bottom thin lines the maximum and minimum values of the distribution (see Gries 2009, 119 on this type of plot). ELF texts use on average 33% of infrequent word combinations, while for NAT texts the percentage is 25%, a statistically very significant difference (ELF $M = 33.56$, $SD = 8.42$, NAT $M = 25.18$, $SD = 11.28$, $t(46) = 3.552$, $p$ (two-tailed) < .01).[13] The analysis was then repeated taking into account only pairs with FQs of 0 or 1 in ukWaC, i.e. pairs which are either unattested or would not count as collocations within a frequency-based paradigm (since a minimum frequency of 2 is required for a word pair to count as a collocation, i.e. a "recurrent" sequence of words). This was done to test the hypothesis that the difference between ELF and NAT texts is due to overuse of word combinations which are *not* collocations in English, rather than *infrequent* collocations in English. Again, the difference was found to be very significant, with ELF texts displaying an average of 24% of "non-collocations" vs. 17% of the NAT texts (ELF $M = 24.85$, $SD = 8.34$, NAT $M = 17.55$, $SD = 10.82$, $t(46) = 2.699$, $p$ (two-tailed) < .01). Cursory browsing of these word combinations revealed that non-collocations are not only more frequent in ELF texts, but also different in "type" from NAT ones: the former include word sequences in a foreign language ("*cette page*", and "*el estudiante*", respectively from a French and Spanish university), misspelled words ("*new aera*", "*environnemental sciences*") and what Nesselhauf (2005, 39) calls "deviant" collocations, i.e. "combinations with questionable acceptability" where one or both elements of the pair can be replaced by more standard/appropriate lexical

items (e.g. "*services enjoyment*", "*innovatory partnerships*", and "*complementary piano*", referring to the name of a compulsory course in Italian music schools, i.e. "*piano complementare*"). Non-collocations in NAT texts tend instead to be combinations which are idiosyncratic to a single university (e.g. "*central-London campus*", "*AntiMal consortium*"), or unusual but intuitively plausible pairs (e.g. "*award-winning facilities*", "*green university*").

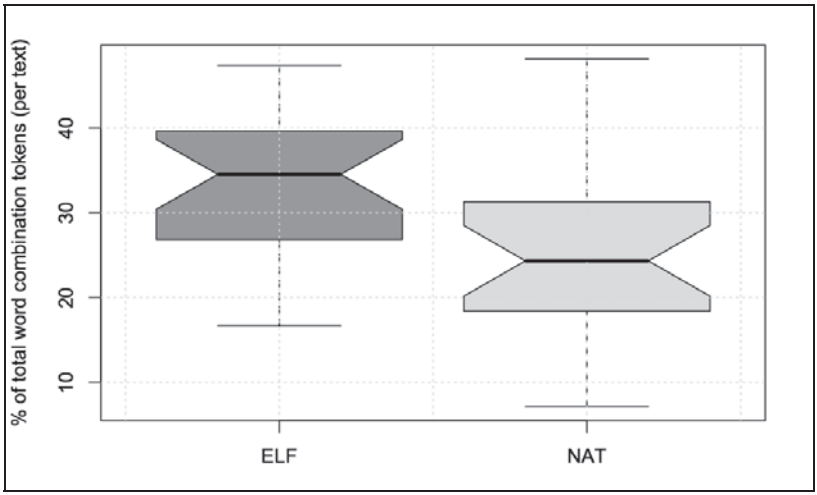



**Figure 1. Percentages of word combinations with FQ < 5.**

Moving on to the second part of the analysis, we consider "strong" collocations according to two lexical association measures. The first of these is the *t*-score: pairs were considered to be "strong" collocations if they obtained a $t$ value $\geq$ 10. Figure 2 represents graphically the percentages of collocation tokens satisfying the condition in the NAT and ELF texts. In this case, NAT texts are found to use strong collocations slightly more than ELF texts, but the difference is not significant (ELF $M$ = 26.38, $SD$ = 8.74, NAT $M$ = 29.41, $SD$ = 10.69, $t(46)$ = -1.0725, $p$ (two-tailed) > .05).

As suggested by Durrant and Schmitt (2009, 171–172), results pertaining to collocation tokens do not provide information as to the degree to which the *same* collocations are repeated in texts, i.e. as to whether these are "characterized by the repeated use of a small repertoire of collocations" (2009, 171). It might be, e.g. that the non-significance of the difference between ELF and NAT percentages of strong collocations

derives from a tendency of the former to repeat a "few favoured formulas", as suggested by other studies on non-native writing (2009, 172). This hypothesis, however, is not supported by the data: even if collocation types are taken into account, the difference between NAT and ELF texts is still not significant (ELF $M$ = 25.70, $SD$ = 8.77, NAT $M$ = 27.87, $SD$ = 9.31, $t$(46) = -0.8298, $p$ (two-tailed) > .05).

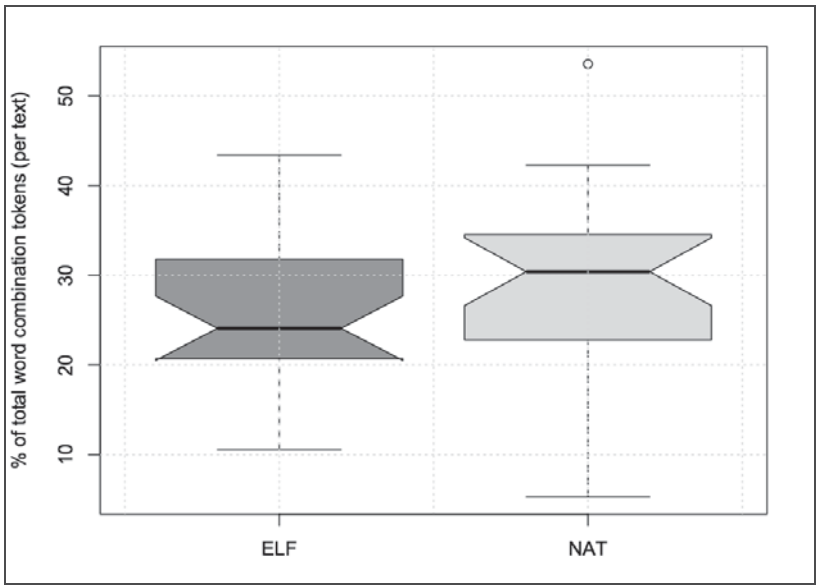



**Figure 2. Percentages of collocations with $t \geq 10$.**

The last analysis, focusing on the use of strong collocations according to the MI score (MI ≥ 7), revealed a greater use of collocations in the NAT texts, this time at a statistically significant level (ELF $M$ = 12.02, $SD$ = 7.65, NAT $M$ = 17.66, $SD$ = 7.63, $t$(46) = -2.555, $p$ (two-tailed) < .05). As is the case with $t$-score results, analysis of collocation types confirms the same trend as that for collocation tokens (ELF $M$ = 12.04, $SD$ = 7.12, NAT $M$ = 17.11, $SD$ = 7.26, $t$(46) = -2.4407, $p$ (two-tailed) < .05), suggesting that repetition of word pairs does not play a greater role in the ELF texts than it does in the NAT ones.

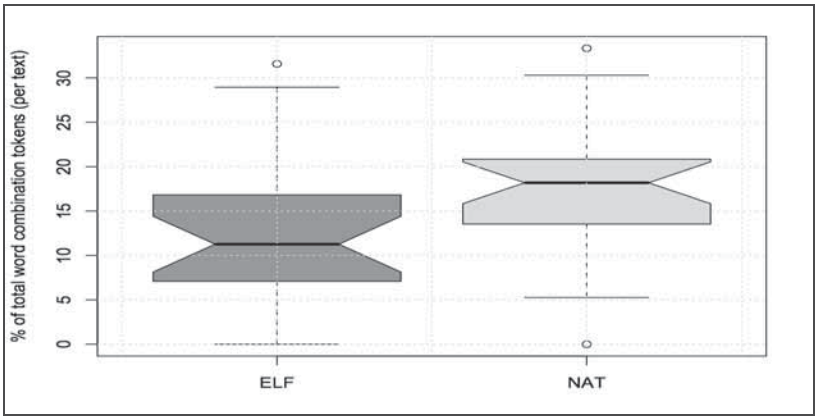



**Figure 3. Percentages of collocations with MI ≥ 7.**

On the whole, these results reveal a somewhat different picture compared to the work by Durrant and Schmitt (2009). First, ELF texts seem *in general* to rely to a lesser extent on phraseological items than their native counterparts, as signalled by their underuse of strong collocations – not always reaching significant levels –, and by the significant overuse of infrequent word combinations, especially of pairs which are unattested in "general English" (as represented by the ukWaC corpus). This is at odds with the "conservatism" observed by the previous study in the production of non-native speakers: ELF texts appear to use more novel combinations, which may or may not be the result of interference from their L1. Second, we could not confirm the trend whereby non-native speakers "over-rely on forms which are […] common in the language" (Durrant and Schmitt 2009, 174): such "over-reliance" could not be observed in the ELF texts taken into account, which display patterns of use of very frequent collocations that are not significantly different from native texts, both in terms of overall number of collocations produced (collocation tokens) and of their variety (collocation types). Third, a significant pattern of underuse by ELF texts is observed in the case of collocations with high MI values, thus reinforcing the hypothesis that these less frequent but strongly associated items are more salient for native than non-native/ELF speakers (2009, 175).

The picture that emerges is one where ELF texts seem to conform only to a limited extent to native norms, and to diverge from them in ways that are slightly different from the non-native texts (mainly produced in the classroom) of Durrant and Schmitt (2009). In particular, the use of

unattested word combinations in ELF web texts – which in a learning environment would most likely be labelled as "errors" –, might hint at either deficient writing skills and/or poor editing (cf. the typos and the word combinations in a foreign language), or at cases in which "ELF generates […] preferences of its own, possibly of those kinds that people from a large number of L1 varieties feel comfortable with" (Mauranen 2010, 19). Similarly, the underuse of highly idiomatic, native-like collocations might derive either from less developed phraseological competence or from

> the specific nature of lingua franca communication, which requires [speakers] to use the linguistic code as directly [i.e. unidiomatically] as possible even if their language proficiency would allow them to sound more native-like than they actually do. (Kecskes 2007, 22)

Assessing whether the observed phraseological deviations are actually the result of inferior language proficiency or rather of more or less conscious strategies aiming to facilitate international communication will make an interesting object for future research.

## 5. Conclusion

Institutional academic genres have been largely neglected as objects of study in EAP research despite their importance for both applied and descriptive purposes. The present contribution has aimed at stimulating interest in the topic by pursuing two aims: first, it has introduced acWaC-EU, a large corpus of institutional academic web texts produced by European universities. Second, it has presented a case study in which acWaC-EU was used to compare the phraseology produced by universities based in countries where English is a native language vs. a lingua franca.

It was argued that ELF university homepages display (phraseological) patterns which are only partially consistent with previous studies of non-native language, and that these deviations might or might not derive from conscious strategies to target an international audience. Future studies should try to shed light on modes of production of institutional academic texts on the web (e.g. are they written by native or non-native speakers? To what extent do they reflect "corporate" communicative strategies rather than preferences of single authors? Who is the targeted population?), which would provide a more precise picture of the parties involved in the communicative exchange. Confirmation of the present results should also be sought in other genres, e.g. academic module descriptions, which, while being as important as homepages, e.g. for attracting international students,

might represent more spontaneous/less edited modes of production on the part of academic staff. Several other ways forward come to mind, e.g. replicating our study, which was conducted on Romance language countries, with countries from other language families; extending the comparison to other aspects of phraseology and other linguistic features beyond phraseology; and, in the longer term, extending the corpus to include universities from countries outside of Europe. By making the tools and heuristics used to build acWaC-EU available to the research community and illustrating the corpus potential, we hope to stimulate the interest of ELF and academic discourse scholars to pursue some of these challenging issues further.

## Notes

[1] http://ec.europa.eu/education/higher-education/bologna_en.htm [Last visited: 21/07/2014]

[2] Fernández Costales (2012) and Candel Mora (2012) also deal with institutional websites, but their focus is on translation into English rather than the linguistic features of this specialised variety of ELF.

[3] http://www.webometrics.info/en/Europe [Last visited: 21/07/2014]

[4] Although interesting, a multimodal analysis of web pages taking into account their visual components is not among our current priorities. For this reason we only keep plain text, which also makes it easier to (index and) consult the corpus with otherwise powerful corpus query tools like the Corpus Workbench.

[5] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/ [Last visited: 21/07/2014]

[6] http://cwb.sourceforge.net/ [Last visited: 21/07/2014]

[7] In particular, the main text type taken into account is the (argumentative) academic essay; in the case of native speakers, essays and opinion articles from newspapers are also included in the corpus (which raises some doubts about the actual comparability of the two subcorpora).

[8] http://www.natcorp.ox.ac.uk/ [Last visited: 21/07/2014]

[9] The original study also aimed at evaluating the method's results for short and long texts, and for this reason it included 96 texts (48 short + 48 long ones).

[10] The corpus is split into 90 million word parts: two of these were randomly selected (ukWaC-1 and ukWaC-11).

[11] http://www.collocations.de/software.html[Last visited: 21/07/2014]

[12] DS carry out a more fine-grained analysis, taking into account different ranges of *t* and MI values; this is not done in this case study to reduce the level of complexity.

[13] Analyses were carried out with the R statistical software (http://www.r-project.org/ [Last visited: 21/07/2014]). All distributions discussed in this section were checked for normality before *t*-tests for independent samples were carried out.

## References

Afros, Elena, and Catherine F. Schryer. 2009. The genre of syllabus in higher education. *Journal of English for Academic Purposes* 8(3): 224-233.

Altbach, Philip G., and Jane Knight 2007. The internationalization of higher education: Motivations and realities. *Journal of Studies in International Education* 11(3-4): 290-305.

Baroni, Marco, and Silvia Bernardini. 2004. BootCaT: Bootstrapping corpora and terms from the web. In *Proceedings of LREC 2004*, Lisbon, Portugal, 1313-1316. ELDA.

Baroni, Marco, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The Wacky Wide Web: A collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation* 43(3): 209-226.

Bernardini, Silvia, Adriano Ferraresi, and Federico Gaspari. 2010. Institutional academic English in the European context: A web-as-corpus approach to comparing native and non-native language. In *Professional English in the European context: The EHEA challenge*, ed. Angéles Linde López and Rosalía Crespo Jiménez, 27-53. Bern: Peter Lang.

Biber, Douglas. 2006. *University language: A corpus-based study of spoken and written registers*. Amsterdam: John Benjamins.

Caiazzo, Luisa. 2011. Hybridization in institutional language: Exploring we in the 'About us' page of university websites. In *Genre(s) on the move: Hybridization and discourse change in specialized communication*, ed. Srikant Sarangi, Vanda Polese and Giuditta Caliendo, 243-260. Naples: Edizioni Scientifiche Italiane.

Callahan, Ewa, and Susan C. Herring. 2012. Language choice on university websites: Longitudinal trends. *International Journal of Communication* 6: 322-355.

Candel-Mora, Miguel A. 2012. Design and exploitation of a corpus of university ECTS course catalogues for the implementation of a hybrid computer-assisted translation workstation. In *Translation studies: Old and new types of translation in theory and practice*, ed. Lew Zybatow, Alena Petrova and Michael Ustaszewski, 105-110. Frankfurt am Main: Peter Lang.

Church, Kenneth W., and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics* 16(1): 22-29.

Connor, Ulla, and Thomas A. Upton. 2004. The genre of grant proposals: A corpus linguistic analysis. In *Discourse in the professions: Perspectives from corpus linguistics*, ed. Ulla Connor and Thomas A. Upton, 235-256. Amsterdam: John Benjamins.

Depraetere, Heidi, Joachim Van den Bogaert, and Joeri Van de Walle. 2011. Bologna translation service: Online translation of course syllabi and study programmes in English. In *Proceedings of the 15th Conference of the European Association for Machine Translation*, ed. Mikel L. Forcada, Heidi Depraetere and Vincent Vandeghinste, 29-34 Leuven, Belgium.

Durrant, Philip, and Norbert Schmitt. 2009. To what extent do native and non-native writers make use of collocations? *International Review of Applied Linguistics in Language Teaching* 47(2):157-177.

Fairclough, Norman. 1993. Critical discourse analysis and the marketization of public discourse: The universities. *Discourse & Society* 4(2): 133-168.

Fernández Costales, Alberto. 2012. The internationalization of institutional websites: The case of universities in the European Union. In *Translation Research Projects 4*, ed. Anthony Pym and David Orrego-Carmona, 51-60. Tarragona: Intercultural Studies Group.

Ferraresi, Adriano, Silvia Bernardini. 2013. The academic web-as-corpus. In *Proceedings of the 8th Web as Corpus Workshop*, ed. Stefan Evert, Egon Stemle and Paul Rayson, 53-62. Lancaster, UK.

Gesuato, Sara. 2011. Course descriptions: Communicative practices of an institutional genre. In *Genre(s) on the move: Hybridization and discourse change in specialized communication*, ed. Srikant Sarangi, Vanda Polese and Giuditta Caliendo, 221-241. Naples: Edizioni Scientifiche Italiane.

Hyland, Ken. 2011. Projecting an academic identity in some reflective genres. *Ibérica: Revista de la Asociación Europea de Lenguas para Fines Específicos* 21: 9-30.

Hyon, Sunny, and Rong Chen. 2004. Beyond the research article: University faculty genres and EAP graduate preparation. *English for Specific Purposes* 23(3): 233-263.

Jenkins, Jennifer. 2011. Accommodating (to) ELF in the international university. *Journal of Pragmatics* 43(4): 926-936.

Kecskes, Istvan. 2007. Formulaic language in English Lingua Franca. In *Explorations in pragmatics: Linguistic, cognitive and intercultural aspects*, ed. Istvan Kecskes and Laurence R. Horn, 191-219. Berlin and New York: Mouton de Gruyter.

Mauranen, Anna. 2010. Features of English as a lingua franca in academia. *Helsinki English Studies* 6: 6-28.

Mautner, Gerlinde. 2005. For-profit discourse in the nonprofit and public sectors. In *Language, communication and the economy*, ed. Guido Erreygers, Geert Jacobs, 25-44. Amsterdam: John Benjamins.

Meunier, Fanny, and Sylviane Granger. eds. 2008. *Phraseology in foreign language learning and teaching*. Amsterdam: Benjamins.

Morrish, Liz, and Helen Sauntson. 2013. "Business-facing motors for economic development": An appraisal analysis of visions and values in the marketised UK university. *Critical Discourse Studies* 10(1): 61-80.

Nesselhauf, Nadja. 2005. *Collocations in a learner corpus*. Amsterdam: John Benjamins.

Pawley, Andrew, and Frances H. Syder. 1983. Two puzzles for linguistic theory: Nativelike selection and nativelike fluency. In *Language and communication*, ed. Jack C. Richards and Richard Schmidt, 191-225. London: Longman.

Römer, Ute. 2010. Establishing the phraseological profile of a text type: The construction of meaning in academic book reviews. *English Text Construction* 3(1): 95-119.

Saichaie, Kem. 2011. *Representation on college and university websites: An approach using critical discourse analysis*. Ph.D. thesis, University of Iowa.

Simpson-Vlach, Rita C., and Sheryl Leicher. 2006. *The MICASE handbook: A resource for users of the Michigan corpus of academic spoken English*. Ann Arbor: The University of Michigan Press.

Sinclair, John McH. 2004. Corpus and text – Basic principles. In *Developing linguistic corpora: A guide to good practice*, ed. Martin Wynne, 1-16. Oxford: Oxbow Books.

Swales, John M. 2004. *Research genres. Explorations and applications*. Cambridge: Cambridge University Press.

Webster, Gary. 2003. Corporate discourse and the academy. A polemic. *Industry & Higher Education* 17(2): 85-90.

Wolter, Brent, and Henrik Gyllstad. 2011. Collocational links in the L2 mental lexicon and the influence of L1 intralexical knowledge. *Applied Linguistics* 32(4): 430-449.

Wray, Alison. 2002. *Formulaic language and the lexicon*. Cambridge: Cambridge University Press.

# CHAPTER TEN

# STUDYING ELF INSTITUTIONAL WEB-BASED COMMUNICATION BY UNIVERSITIES: COMPARISON AND CONTRAST WITH ENGLISH NATIVE TEXTS

GIUSEPPE PALUMBO
UNIVERSITÀ DI TRIESTE, ITALY

## 1. Introduction

Higher education institutions are increasingly competing for both students and staff in a global marketplace. The collaborative research networks they are part of are also increasingly established at an international level. University websites have come to reflect this international dimension: they may cater for the existing population of international students or have a more overtly advertising function as they try to attract prospective students from abroad. In non-English speaking countries, this often means that a university website must provide information in more than one language and that English is often chosen as the lingua franca used to address the international audience. The choice of languages and the type of information provided may also be influenced by the presence of international exchange programmes for students and staff, such as the Erasmus programme within the EU or the Marco Polo programme between Italy and China.

Designing a multilingual university website is subject to various considerations and constraints to do with the intended purpose of the website and the influence and requirements of the university "stakeholders", i.e. students and parents, teachers and researchers, administrators, funding institutions and the government. More specific questions to consider when designing multilingual material for a university website include: the basic design of the website, i.e. whether content and

services should be provided equally in all designated languages or whether different content should be given in different languages (through ad hoc creation or translation, or both); the number of languages; the awareness of cultural issues in presenting material to an international audience; the availability of resources (in terms of manpower and translation tools); technical issues such as the interaction with existing content management systems and the maintenance of multilingual materials (e.g. in the form of translation memories); a measurement of the effectiveness of the material provided in other languages.

An additional factor to take into account is the increasing diversification of the information and services provided on-line by universities: most university websites today act as portals for a wide range of on-line transactions, including enrolment, course delivery and support, and library lending and research. All or some of these transactions may also need to be conducted in a foreign language, which means that a multilingual dimension is also required for the administrative functions of the university. A considerable amount of translation or foreign-language drafting is carried out both for the general public and 'behind the scenes', as the administrative and departmental units of a university exchange international agreements and student certificates drawn up in more than one language. This in turn gives rise to the establishment of accepted terminological equivalents for degree programme denominations, titles and administrative procedures.

The present study is an attempt at looking at ELF texts produced by universities as examples of hybrids that can be characterized with respect to their counterparts produced in national, native-English contexts. The approach adopted in the analysis is the same as that proposed in the collection of studies reported on in Ondelli (2013), where, however, the focus was somewhat different: in those studies the aim was that of characterizing the hybrid features of texts emerging in multilingual contexts, such as the EU institutions, and the influence they may be exerting on comparable texts produced at national level. In the present study, the analysis will look at some features of comparable sets of texts written in ELF and in two national varieties of English with a view to constructing their respective profiles at the morpho-syntactic level and relating this profile to the way the texts realize their main, shared function. Employing basic corpus-linguistic methods, the analysis does not aim at exhaustiveness but is meant to elucidate some general features of the texts that could later be subjected to a more in-depth study, possibly on a larger corpus.

## 2. English as a lingua franca at university

The status of English as the lingua franca of academics is undisputed. The research on the English used by academics has traditionally concerned the textual and discursive practices involved in the negotiation of scientific knowledge and especially on the written reporting of research as embodied in research articles (Bhatia 1993; Swales 1993), poster presentations (D'Angelo 2011), abstracts (Bondi 2004), book review articles (Diani 2012) and other research genres. Academics, however, are involved in various other communicative practices, a concise but useful overview of which is given in Gesuato (2011). The list includes "participation in forums [...] for sharing knowledge", such as conferences, workshops, presentations and lecture notes; the "handling of professional and social relationships with peers, superior and learners", e.g. reports and reference letters; and finally "administrative tasks", e.g. writing minutes and filling out registers (Gesuato 2011, 221-222).

In quite a few of these practices, non English-speaking academics are today increasingly required to use English, often to communicate to other non-native speakers. The label "English as a lingua franca", or ELF, is based on the definition of lingua franca as "a vehicular language used by speakers who do not share a first language" (Mauranen 2012, 8). Research on English in academic settings with explicit reference to the notion of ELF has emerged as a recent tradition of research that has reversed the usual order of linguistic research, concentrating on spoken modes before written modes (see, for instance the ELFA project, described in Mauranen et al. 2010). Even in ELF writing, however, new, hybrid forms are appearing, often determined by the newer modes of communication afforded by web-based and online environments. Whereas this hybridity may be more pronounced in web-based genres relying on interaction (e.g. blogs and forums), hybrid traits could also be hypothesized for other web-based genres of a more static nature but increasingly produced through editorial processes that see no involvement, or only a marginal involvement, of English native speakers. To remain in the academic setting, recent studies of non-research written genres include Gesuato's (2011) own investigation of academic course descriptions and Bernardini et al.'s (2010) analysis of academic programme descriptions. University websites as a special case of institutional communication are investigated by Caiazzo (2010, 2011), who points out how the web has become a particularly suitable communicative environment for the mixing of informative and promotional functions already noted in the pre-web era by Fairclough (1993), who, from the angle of critical discourse analysis,

charted the changes produced within British universities in genres such as undergraduate prospectuses and job ads.

## 3. ELF, hybridization, translation

Universities all over the world today produce a growing body of written ELF materials for purposes of institutional communication, online interaction with students, and collaboration with other universities on teaching and research programmes. The wide range of ELF documents includes promotional web pages and brochures, press releases, study guides, student regulations, and various types of contracts and agreements between universities. The influence of Anglocentric textual models for the production of some of these documents, especially those having a more overtly promotional function, has been characterized by a tendency to "hybridise local identities" (Gotti 2007, 145). ELF can be seen as "a dynamic and hybrid language whose complexity cannot be fully grasped without taking into account its interaction with other languages and cultures" (Taviano 2013, 156). This can be observed at various levels of description, the most obvious of which is perhaps terminology. In spoken interactions, ELF users "typically find themselves in situations where discourse norms are not clear or given: group norms are negotiated within ELF groups by participants, none of whom can claim the status of a linguistic model" (Mauranen 2012, 7). This may be different in written communication, where it is more likely that ELF users decide to take a national variety as their model, either deliberately or because they have been instructed to do so as part of the editing cycles that a text destined for publication normally goes through. Other factors, however, may guide or constrain text production, with varying degrees of awareness on the part of language users.

Endonormativity within a community of ELF users at written level may be more apparent at the lexical or terminological level. In higher education this is what happened with the vocabulary of international student mobility within the EU, agreed upon, in English, by bodies in which native speakers of English are often a minority. In cases like these, the best solutions "need not be the most standard-like or native-like" (Mauranen 2012, 8) but may arise out of non-language-related compromises between the parties involved (e.g. a deliberate willingness to remain vague) or for the interference from another language (as is often the case with French in EU institutions).

The consideration of international scenarios in the discussion of how texts are produced, or translated, has often been linked to the question of

hybridity. Translations, in particular, have often been presented as the quintessentially hybrid texts in that they often display "features that somehow seem 'out of place'/'strange'/'unusual' for the receiving culture, i.e. the target culture" (Schäffner and Adab 2001, 169). In particular, translations are sometimes described as texts showing a significant degree of markedness with respect to texts produced by native speakers of the target language, mainly as a result of either the overrepresentation of certain traits or features (e.g. a higher frequency of occurrence of certain items or patterns) or the underrepresentation, or even the absence, of other traits that are frequently found in native texts. Qualifying translations as hybrids, however, may lead us to ignore comparisons with other, possibly more hybrid, modes of text production. As Pym (2001, 203) points out, there are various contexts in which "sources are becoming more hybrid than their translations". Typically, in such contexts texts producers are not translators but inhabit the same "intercultural space" as translators: people who use a foreign language (very often ELF) for interacting with each other and producing drafts and official documents. The texts produced by the EU institutions, for instance, may well be characterized as hybrids with respect to comparable texts produced in national contexts.

The parallelism between ELF and translated language may be considered questionable from a theoretical point of view, and indeed various ELF researchers have often resisted proposing it. According to Mauranen (2012), ELF is more appropriately defined as an instance of "second order language contact", or the site of contact between different hybrids: in a typical ELF situation, and especially in spoken interaction in a group of speakers, "a large number of languages are each in contact with English, and it is these contact varieties (similects) that are, in turn, in contact with each other" (Mauranen 2012, 30). "First order language contact", on the other hand, is to be identified with situations in which speakers of two different languages use of one of them to communicate with each other. On the basis of Mauranen's distinction, a book or a newspaper article translated into English for a native-English audience could be equated to "first order language contact": the source-text author adopts (or is made to adopt by the translator) the language of the target-text reader. In discussing the distinction between first and second order contact Mauranen (2012, 28-29) herself does not mention translation but second language acquisition as a case of first order contact: the L1 of different learners of English gives rise to varieties (the above-mentioned "similects", also popularly known as Spanglish, Swinglish, Finglish and the like), each of which shares a set of transfer features closely associated with the specific characteristics of the corresponding L1. In short, ELF is

not to be confused with any one of these similects but is shaped by the interaction of various similects.

For the purposes of the present study, it is argued that when translation into English or drafting in English is addressed at an international audience (which is more often than not constituted by a majority of non-native speakers) the kind of contact established by English shares some features of the settings in which English is used in spoken interactions between non-native speakers, where communicative effectiveness is the overriding aim. In such scenarios, whether translations are carried out by native or non-native speakers may be considered irrelevant (House 2013).

As pointed out by Mauranen (2012, 22-23) herself, a vital aspect of the conceptual models used to describe the social groupings that typically use ELF is multilingualism. Users of ELF often operate and interact with each other in multilingual environments where they may employ different linguistic repertoires with the overriding aim of communicating effectively. This circumstance tends to be ignored in social models of language, normally and tacitly constructed on the basis of monolingual views of language use. Yet, the description of the ways in which norms evolve or emerge or effective communication is negotiated in ELF situations would probably benefit from a consideration of the multiple codes that language users may have at their disposal and the ways these are deployed in or constrain the communication. In this respect, ELF research might be seen to profit from the insights and methods of the other domains where language use is characterized against the backdrop of language contact and multilingual repertoires, that is, translation studies and contrastive linguistics.

## 4. Data description

The small corpus compiled for the present case study aims to explore differences and similarities in three sets of web-based texts published by European, British and North-American (i.e. US and Canadian) universities, with European universities taken to represent the ELF group. The texts are all taken from the institutional websites of the universities and, more specifically, from the sections explicitly addressed at an audience of "international students". For the ELF group, the texts are taken from the websites of European universities with a strong international vocation: the selected websites were chosen among those that have received the "ECTS label", awarded to Higher Education Institutions who successfully apply the principles of the European Credit Transfer and Accumulation System (ECTS). In order to be awarded the ECTS label, an

institution must publish both an Information Package and a Course Catalogue in English on its website, with the Information Package expected to offer detailed descriptions of study programmes, units of learning, university regulations and student services. The present corpus was collected by downloading the ECTS Information Packages of 13 universities, totalling 249,159 tokens.

The other two corpus components were collected so as to provide a double comparative element for the analysis of the EU universities corpus (from now on abbreviated to UniEU), based on the two major national varieties of English, or the varieties that drafters and editors of the English-language sections of European websites are more likely to consider as providing a standard for reference. The British component (UniUK) includes text downloaded from the "International" section in the websites of 10 British universities, for a total of 303,755 tokens. The North-American component (UniNAm) collects text from the same section of the websites of 9 US universities and 1 Canadian university, for a total of 456,732 tokens. The three corpus components are not balanced in terms of size, which is due to the fact that all of the British and North-American universities considered for inclusion in the corpus tended to provide a higher amount of information in the International section than the selected European universities (whose number was expressely brought up from the original 10 to 13 so as to increase the amount of text available for analysis). In the analysis, normalization of frequencies and other methods will be employed to ensure comparability of results between the corpus components. It should be noted, in any case, that the size of the corpus is too small to allow statistically significant conclusions.

## 5. Results and discussion

The analysis aims at identifying some particular features of English used in non-native contexts against the backdrop of native-English texts produced in similar settings, and published on the same medium. The rationale for the study is that ELF written texts can be investigated along lines that are similar to those employed in studies of translated language that aim at uncovering the distinctive features of translated texts as opposed to non-translated texts. In particular, the three corpus components have been compared for the following aspects: distribution of part-of-speech categories; distribution of verb tenses; ratio of modal verbs to non-modal verbs; use of pronouns. The assumption is that these aspects may give an indication of the possible differences and similarities in how the

three sets of texts realize their common, predominant orientational-directive function from the point of view of morpho-syntax.

In more general terms, the attempt made by the present case study is to try and establish a method of analysis that could be employed to complement more focused analyses of lexico-grammatical features based on frequencies of individual items and multi-word units. ELF texts are likely to differ from native texts along the *paradigmatic* axis because of interference from the speakers or writers' L1 or due to the specialised nature of domain in which they are produced, which gives rise to specialised terminology or domain- or genre-specific phraseologisms and lexical bundles. Such differences (or similarities) may be relatively easily identified through corpus searches and frequency counts. The present analysis looks at possible differences across sets of texts along the *syntagmatic* axis, focusing on the elements that can be more easily investigated by means of a part-of-speech tagged corpus. Other syntactic aspects (such as sentence length and sentence structure) fall outside the scope of the present study but their investigation would certainly help in completing the picture emerging from the analysis illustrated here.

## 5.1. Distribution of part-of-speech categories

The three corpus components were analyzed using the part-of-speech (POS) tagger now available in the Sketch Engine (Kilgarriff et al. 2004). The distribution of POS categories in each corpus component is shown in Table 1. Please note that in the table the total number of tokens for each corpus component is different from that given above, in that only the total number of tokens actually considered by the automatic POS tagger is given. (Also note that the percentages given in the table do not add up to 100% as some categories deemed not relevant for the analysis were left out, e.g. cardinal numbers, interjections and foreign words).

As the table shows, of the three corpus components UniEU is the one with the highest proportion of nominal forms, as regards both common and proper nouns. This may be due to the influence of some of the L1, or source languages, 'covertly' represented in the corpus. Romance languages, in particular, are generally presented as using nominalised forms more frequently than English, which may be the origin for the higher percentage of nouns in UniUE in comparison with both UniUK and UniNAm. In UniEU, moreover, the higher percentage of nouns is also perhaps linked to the higher percentage of both determiners, especially articles, and adjectives.

| | UniEU (N= 249159) | | UniUK (N=303755) | | UniNAm (N=456732) | |
|---|---|---|---|---|---|---|
| Nouns | 60149 | 24.14% | 66807 | 21.99% | 101048 | 22.12% |
| Proper nouns | 30196 | 12.12% | 944 | 4.41% | 687 | 3.20% |
| Verbs | 29962 | 12.02% | 40846 | 13.45% | 59252 | 12.97% |
| Modals | 3761 | 1.51% | 5965 | 1.96% | 8091 | 1.77% |
| Adjectives | 18362 | 7.37% | 21343 | 7.07% | 30873 | 6.76% |
| Adverbs | 7076 | 2.84% | 8235 | 2.71% | 11180 | 2.45% |
| WH-adverbs | 651 | 0.26% | 1021 | 0.33% | 1434 | 0.31% |
| Determiners | 21546 | 8.65% | 22718 | 7.48% | 34183 | 7.48% |
| WH-determiners | 846 | 0.33% | 937 | 0.31% | 1591 | 0.35% |
| Prepositions and subordinating conjunctions | 30706 | 12.32% | 35174 | 11.56% | 50134 | 10.98% |
| Pronouns | 8820 | 3.54% | 14128 | 4.65% | 17206 | 3.78% |
| WP-pronouns | 613 | 0.25% | 1359 | 0.45% | 1656 | 0.36% |
| Coordinating Conjunctions | 8850 | 3.55% | 11523 | 3.79% | 18342 | 4.01% |

**Table 1. Distribution of part-of-speech categories in the three corpus components.**

With respect to nouns, and particularly to noun-headed phrases, a way of checking for the possible influence on L1 languages is to see how pre- and postmodification in such phrases (e.g. *course of study* vs *study programme*; for a discussion, see Biber et al. 2001, 578-602) are distributed in each set of texts, the assumption being that in the ELF set the preference for postmodified structures is more marked than in the native sets or a possible interference of some L1s. Rough measures in this respect can be obtained by counting occurrences of noun sequences with and without intervening prepositions and considering them as proportions of the total number of nouns in a given set of texts. Table 2 shows results for such counts of the most frequent postmodified structure in all three sets of texts (NOUN + *of/of the* + NOUN), compared with premodified noun-headed phrases using one or two nouns or a genitive suffix.

| | UniEU | | UniUK | | UniNAm | |
|---|---|---|---|---|---|---|
| NOUN + of (the) + NOUN | 3897 | 4.31% | 4021 | 3.76% | 5857 | 3.38% |
| NOUN + NOUN | 10362 | 11.47% | 10856 | 10.16% | 16742 | 9.65% |
| NOUN + NOUN + NOUN | 1146 | 1.27% | 1458 | 1.36% | 2088 | 1.20% |
| NOUN + 's + NOUN | 641 | 0.71% | 574 | 0.54% | 1527 | 0.88% |
| Total nouns | 90345 | | 106881 | | 173419 | |

**Table 2. Frequencies of some noun sequences in the three corpus components (with an indication of their proportion, in %, to the total number of nouns).**

As expected, the ELF set (UniEU) does use a higher proportion of premodified noun phrases but, interestingly, it also makes use of more NOUN+NOUN sequences than the other two sets.[1] Premodification through a genitive suffix is also more frequent in the ELF set than in one of the native sets (UniUK). To sum up, whereas the ELF set shows a certain preference for nominal over verbal structures, when the particular profile of the nominal structures is compared across the three corpus components, the ELF set does not seem to be in stark contrast with the two native sets concerning the distribution of pre- and postmodified structures.

As regards verbal forms, both UniUK and UniNAm (the native corpora) show higher proportions than UniEU, which again may be due to a covert influence of some L1 languages represented in the corpus. In particular, the corpus was analyzed so as to establish whether the distribution of tenses showed particular patterns of variation across the three components, with reference to both non-progressive and progressive tenses. The focus on progressive forms was suggested by the findings of studies on spoken ELF interactions (e.g. Ranta 2006) indicating that speakers of ELF tend to make a more "extended use" of the progressive than native speakers, regardless of their L1. The breakdown of verb tenses in the indicative is given in Table 3. (The other verbal moods, i.e. the conditional and the subjunctive, were left out of the analysis as they only play a minor role in the corpus, perhaps due to the fundamentally informative nature of the texts under analysis).

| | UniEU (N= 249159) | | UniUK (N=303755) | | UniNAm (N=456732) | |
|---|---|---|---|---|---|---|
| Simple Present | 10016 | 4.020% | 13180 | 4.339% | 18342 | 4.016% |
| Simple Past | 950 | 0.381% | 899 | 0.296% | 1624 | 0.356% |
| Future (with *will*) | 1094 | 0.439% | 2022 | 0.666% | 2308 | 0.505% |
| Present Perfect | 487 | 0.195% | 785 | 0.258% | 876 | 0.192% |
| Past Perfect | 18 | 0.007% | 6 | 0.002% | 4 | 0.001% |
| Future Perfect | 3 | 0.001% | 7 | 0.002% | 5 | 0.001% |
| Total non-progressives | 12568 | 5.044% | 16899 | 5.563% | 23159 | 5.071% |
| | | | | | | |
| Present Progressive | 158 | 0.063% | 418 | 0.138% | 444 | 0.097% |
| Past Progressive | 2 | 0.001% | 2 | 0.001% | 6 | 0.001% |
| Future Progressive | 11 | 0.004% | 37 | 0.012% | 21 | 0.005% |
| Present Perfect Progressive | 11 | 0.004% | 19 | 0.006% | 4 | 0.001% |
| Future Perfect Progressive | 0 | 0 | 1 | 0.0003% | 0 | 0 |
| Total progressives | 181 | 0.073% | 487 | 0.160% | 475 | 0.104 |

**Table 3. Frequencies of non-progressive and progressive indicative tenses in the three corpus components (with an indication of their proportion, in %, to the total of tokens in the corpus).**

As the table shows, the progressive is more represented in the native texts (UniUK and UniNAm). More specifically, the tense that is proportionately more frequently represented in both sets is the present progressive, possibly because the native texts prefer it as an unmarked from of the present tense. Contrary to findings for spoken ELF interactions, then, the use of the progressive does not appear to be extended in the analyzed set of written ELF texts. If anything, as seems to be suggested by the results for the present progressive, the ELF texts under analysis may even be said to point to a restricted use of the progressive in comparison with native texts.

One final note about verb usage in the analyzed sets has to do with the future tense. Among the non-progressive verbs, both UniUK and UniNAm display a higher proportion of verbs in the future constructed with *will*, which is explained by the particularly high frequency of the structure *will*

*be + PAST PARTICIPLE* in both sets (very often having *you* as the subject). Particularly frequent examples include *you will be expected* and *you will be required* – ways of expressing directives that are comparatively less frequent in the ELF texts.

## 5.2. Ratio of modals to non-modal verbs

As we have seen, UniEU has lower counts of both modals and other verbal forms than the two native sets. Modals, however, can be looked at in the three corpora not only considering their frequency in relation to all tokens (as in Table 1) but also in relation to the total number of finite verb phrases contained in a set of texts, so as to see whether – in using verbal forms in general – one of the three sets displays a more pronounced preference for modals over other verbs. Table 4 shows the frequencies of finite verb phrases and modals in the three corpus components and indicates what the percentage of modals relative to the total of verb phrases is. (Note that for this count the modal *will* has not been considered within the group of modals but as part of finite verb phrases in the future tense.)

| | UniEU | UniUK | UniNAm |
|---|---|---|---|
| Modals (except *will*) | 2439 | 3438 | 5108 |
| Finite verb phrases | 12568 | 16899 | 23159 |
| Total verbs | 15007 | 20337 | 28267 |
| Percentage of modals | 16.25 | 16.90 | 18.07 |

**Table 4. Frequencies of modals and finite verb phrases in the three corpus components.**

When modals are looked at in this way, they again appear to be less represented in UniUE, although to a lesser extent than in the count given in Table 1. Although the difference in percentages is not such that it clearly sets the non-native texts apart from the native texts, it may still point to the use of modals as a preferred way of realizing the orientational-directive function in the native texts, as opposed to a greater use of alternative constructions for expressing the same function in the non-native texts. The pattern *to be obliged to*, for instance, has 48 occurrences in UniEU, as against only 3 in UniUK and none at all in UniNAm.

## 5.3. Use of pronouns

The higher percentage of personal pronouns in both UniUK and UniNAm as shown in Table 1 would seem to give support to the hypothesis that English language websites written in native settings use a more direct style, i.e. one which more frequently uses pronouns to either address readers directly ("you") or for self-mentions ("we").

A quick way to check this hypothesis is to query the corpus for occurrences of *you +[modal verb]*, comparing them with the frequency of a corresponding more impersonal phrase: *STUDENT + [modal verb]*,[2] which gives the results presented in Table 5.

| | UniEU | UniUK | UniNAm |
|---|---|---|---|
| you + MODAL | 2.79 | 3.48 | 2.37 |
| STUDENT + MODAL | 0.99 | 0.80 | 1.35 |

**Table 5. Relative frequencies (in %), of *you + [modal verb]* and *STUDENT + [modal verb]* in the three corpus components.**

As the table shows, only one of the two native sets of texts decidedly prefers the more direct form (i.e. the personal pronoun) in combination with a modal verb as a form of address to readers. What is more, contrary to the initial assumption, it is not the ELF corpus but the other native component (UniNAm) that more frequently employs the more impersonal pattern *STUDENT + [modal verb]*. Another way of checking the preference for a more direct and personalized style is to see how the universities represented in the corpus refer to themselves, i.e. whether they prefer using the personal pronoun *we* (Table 6) as against other possible, more impersonal phrases such as *the university* or *our university*.

| | UniEU | UniUK | UniNAm |
|---|---|---|---|
| we | 0.14 | 0.41 | 0.19 |

**Table 6. Relative frequency (in %) of the pronoun *we* in the three corpus components.**

Whereas the use of *we* for self-mentions is common in all three sets of texts (but especially so in UniUK), the phrase *the university* is very rarely used in all three corpus components. As the subject in a clause, it only occurs 18 times in UniEU, 8 in UniNAm and 2 in UniUK. The phrase *our university* is never used as the subject in a clause; it occurs, for instance,

only 8 times in UniEU and always as part of more extended nominal groups not functioning as subjects, e.g. *research work at our university* or *the structure and development plan of our university*. To sum up, corpus analysis confirms that all three sets of texts tend to resort to personalized forms, both to address readers and, to an even greater extent, for self-mentions. The results are in line with what was found by Caiazzo (2011) on the use of "we" in British and Indian universities websites.

## 6. Summary and conclusions

The case study presented above, concentrating on some specific morpho-syntactic aspects observed across three different sets of texts, was based on the assumption that the non-native ELF texts would present differences from comparable native texts in the way they realized the orientational-directive function common to both sets of texts. The results have shown, however, that such expected differences are not as pronounced as was expected and that ELF texts are remarkably close to native texts in the way they deploy grammatical resources to realize their predominantly informative function.

Some differences between the ELF set and the two native sets have been observed as regards the use of verbs. Verb forms in general appeared to be less frequent in the ELF set than in the native sets, which was explained in terms of the possible influence of some of the L1 indirectly represented in the ELF corpus (especially the Romance languages). This explanation, however, would need to be confirmed by looking at whether and how the general distribution of nominal and verbal forms varies across the subcomponents of the ELF set. Another feature shared by the two native-English sets is their more frequent use of the progressive, especially in the present tense, which could be a reflection of its use as the unmarked choice in everyday language.

Besides the above differences, however, the analysis of the three corpus components also showed that in a number of aspects the three sets of texts presented a remarkably similar profile. In particular, the similarity concerned not only purely structural aspects (such as the generalised tendency to use premodification in noun phrases) but also the use of patterns pointing to the adoption of similar signals of stance or engagement, such as the heavy personalization of the discourse through the use of pronouns ("we"/"you" as opposed to "the university"/"students"), even in a communicative context that generally refrains from adopting the accented promotional overtones noted in other investigations of web-based institutional communication by universities.

Overall, the analysis presented here seems to point to a certain homogeneity between the non-native and native sets with regard to their structural make-up. In this sense the study would appear to complement the homogeneity emerged elsewhere at the level of rhetorical models in written ELF communication, especially in institutional and corporate web-based genres (Caiazzo 2010; Poppi 2013). The distance between native and non-native varieties emerging from the present study is much shorter than that observed in studies of spoken ELF – an obvious finding in light of the particular character of spoken face-to-face interactions and the overriding role that communicative effectiveness plays in them. The convergence of web-based ELF texts on certain rhetorical models, however, might also be seen as an attempt at guaranteeing communicative effectiveness through the adoption of models judged (more or less intuitively) to be more suitable to the specificities of the medium – an idea worth investigating further but falling outside the scope of the present work.

## Notes

[1] From a methodological point of view, it should be noted that the count of NOUN+NOUN sequences presented in Table 2 includes occurrences of any pair of juxtaposed nouns, meaning that many of the counted sequences may not be actual noun phrases. In this respect, the count of NOUN+NOUN+NOUN sequences is more likely to represent genuine noun phrases, given the lower probability of three nouns occurring one after the other outside of a noun phrase. The ELF set shows no marked difference from the native sets on both counts, suggesting a tendency to premodified structures.

[2] Small capitals indicate a lemma.

## References

Bernardini, Silvia, Adriano Ferraresi, and Federico Gaspari. 2010. Institutional academic English in the European context: A web as corpus approach to comparing native and non-native language. In *Professional English in the European context: The EHEA challenge*, ed. Ángeles Linde López and Rosalia Jiménez Crespo, 27-53. Bern: Peter Lang.

Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad, and Edward Finegan. 1999. *Longman grammar of spoken and written English*. Harlow: Longman.

Bhatia, Vijay K. 1993. *Analysing genre: Language use in professional settings*. London/New York: Longman.

Bondi, Marina. 2004. The discourse function of contrastive connectors in academic abstracts. In *Discourse patterns in spoken and written corpora*, ed. Karin Aijmer and Anna-Brita Stenström, 139-156. Amsterdam/Philadelphia: John Benjamins.

Caiazzo, Luisa. 2010. The "promotional" English(es) of university websites. In *Discourses, communities and global Englishes*, ed. Roberto Cagliero and Jennifer Jenkins, 43-60. Bern: Peter Lang.

—. 2011. Hybridization in institutional language: Exploring we in the "About us" page of university websites. In *Genre(s) on the move. Hybridization and discourse change in specialized communication*, ed. Srikant Sarangi, Vanda Polese and Giuditta Caliendo, 243-260. Naples: Edizioni Scientifiche Italiane.

D'Angelo, Larissa. 2011. The academic poster presentation: An exploration of the genre. In *Genre(s) on the move. Hybridization and discourse change in specialized communication*, ed. Srikant Sarangi, Vanda Polese and Giuditta Caliendo, 191-203. Naples: Edizioni Scientifiche Italiane.

Diani, Giuliana. 2012. *Reviewing academic research in the disciplines: Insights into the book review article in English*. Rome: Officina Edizioni.

Fairclough, Norman. 1993. Critical discourse analysis and the marketization of public discourse: The universities. *Discourse and Society* 4(2): 133-168.

Gesuato, Sara. 2011. Course descriptions: Communicative practices of an institutional genre. In *Genre(s) on the move. Hybridization and discourse change in specialized communication*, ed. Srikant Sarangi, Vanda Polese and Giuditta Caliendo, 221-241. Naples: Edizioni Scientifiche Italiane.

Gotti, Maurizio. 2007. Globalisation and discursive changes in specialised contexts. In *Discourse and contemporary social change*, ed. Norman Fairclough, Giuseppina Cortese and Patrizia Ardizzone, 143-172. Bern: Peter Lang.

House, Juliane. 2013. English as a lingua franca and translation. *The Interpreter and Translator Trainer* 7(2): 279-298.

Kilgarriff, Adam, Pavel Rychly, Pavel Smrz, and David Tugwell. 2004. The Sketch Engine. In *Proceedings of the Eleventh Euralex Congress*, ed. Geoffrey Williams and Sandra Vessier, 105-116. Lorient, France: UBS.

Mauranen, Anna. 2012. *Exploring ELF: Academic English shaped by non-native speakers*. Cambridge: Cambridge University Press.

Mauranen, Anna, Niina Hynninen, and Elina Ranta. 2010. English as an academic lingua franca: the ELFA project. *English for Specific Purposes* 29(3): 183-90.

Ondelli, Stefano (a cura di). 2013. *Realizzazioni testuali ibride in contesto europeo*. Trieste: EUT.

Poppi, Franca. 2013. *Global interactions in English as a lingua franca*. Bern: Peter Lang.

Pym, Anthony. 2001. Against praise of hybridity. *Across Languages and Cultures* 2(2): 195-206.

Ranta, Elina. 2006. The 'attractive' progressive – Why use the -ing form in English as a lingua franca? *Nordic Journal of English Studies* 5(2): 95-116.

Schäffner, Christina, and Beverly Adab. 2001. The idea of the hybrid text in translation: Contact as conflict. *Across Languages and Cultures* 2(2): 167-180.

Taviano, Stefania. 2013. English as a lingua franca and translation. Implications for translator and interpreter education. *The Interpreter and Translator Trainer* 7(2): 155-167.

# PART IV

# PEDAGOGICAL IMPLICATIONS IN EAP

# CHAPTER ELEVEN

# GENRE, CORPUS AND DISCOURSE: ENRICHING EAP PEDAGOGY

MAGGIE CHARLES
UNIVERSITY OF OXFORD, UK

## 1. Introduction

In recent years, there have been many descriptions of the direct use of corpora in English for Academic Purposes (EAP) pedagogy. Reviews by Flowerdew (2009) and Yoon (2011) and the volumes published as follow-up to the highly successful *Teaching and Language Corpora* conferences (e.g. Frankenberg-Garcia et al. 2011; Kübler 2011; Thomas and Boulton 2012) all bear witness to increasing interest and activity.

Building on the groundbreaking work of Johns (1991a, 1991b), corpora have been most widely used in the teaching of written academic discourse at tertiary level. Most of this work has reported on courses in English for General Academic Purposes (Thurstun and Candlin 1998; Kaltenböck and Mehlmauer-Larcher 2005; Yoon 2008; Boulton 2010), although another major area in which direct corpus methods have been used is in courses for translation and English language (Bernardini 2000, 2002; Frankenberg-Garcia 2005; Cresswell 2007; Estling Vannestål and Lindquist 2007; Granath 2009). Corpus-based courses in specific disciplines have been described, including engineering (Mudraya 2006), law (Hafner and Candlin 2007) and business studies (Flowerdew 2012), while Lee and Swales (2006) have reported on the use of self-compiled corpora, which allows for greater individualisation within classes of multi-disciplinary learners.

However, it is noticeable that most, if not all, of the above accounts focus on teaching lexico-grammatical features of writing, with much less attention paid either to genre or discourse issues. To a certain extent, then, the criticism of Swales (2002) that corpus materials present a bottom-up

approach to a "fragmented world" still has some validity. However, although Flowerdew's repeated calls for a more discourse-based approach (1998, 2005) have not yet been fully implemented, some headway has been made in using corpora to teach genres. Thus Bondi (2001) describes a corpus-based procedure and materials for examining research article (RA) abstracts in economics and Gavioli (2005) deals with the same genre in both economics and medicine. Aston (1995) also addresses the discipline of medicine, giving an account of teaching the part-genres of introductions and methods in RAs, while Weber's (2001) work takes the legal essay as his target genre. Following a similar approach to Lee and Swales (2006), in which students build their own corpora, Cortes (2007, 2011) focuses on each part-genre of the RA in turn. She uses readings from applied linguistics, along with students' investigations of their own corpora in order to teach this genre and to raise awareness of cross-disciplinary variation within it.

Like Cortes (2007, 2011), Bianchi and Pazzaglia (2007) have also shown how students can take an active part not only in the investigation, but also in the construction of corpora. In their course on writing a psychology RA, they ask students to contribute an article of their own choosing to the corpus and to analyse and mark it up according to a set of generic moves provided by the teacher-researchers. The advantages of this approach are that the students not only have a specialist corpus to work with, but have invested their own time and effort into helping to build it. This means that they can experience a sense of ownership of the resource, which is lacking when students use corpora constructed by others. It also enables them to understand much better how a corpus is built and the types of decisions that lie behind genre descriptions. I would suggest that this background knowledge is useful in informing their subsequent use of this, and indeed other, corpora and corpus-based reference materials.

One criticism that has been put forward by those opposed to the teaching of genres in classrooms is that such teaching can be rigid and prescriptive (see e.g. Freedman, 1993). It may consist in providing models which are followed blindly, without due attention to the fluidity and variability of genres as socially situated practices arising out of shared communicative purposes. I would argue that the use of corpora in teaching genres is likely to lead to a less prescriptive and more investigative approach to learning genre. If we take the "family resemblance" approach to genres (see Swales 1990, 49), then corpora can be used specifically to problematise genre descriptions and thus to widen the focus from a single prototypical exemplar to a range of less conventionalised, though still acceptable, members of the group. Corpora are ideally suited to this kind

of work because they can provide the evidence and the multiple examples that are necessary to encourage students to take a critical and enquiring approach to genres.

Work on corpus-based genre pedagogy, then, is highly promising, though not yet extensive, and there have also been reports on the teaching of discourse patterns or functions. Flowerdew (2008) examines the problem-solution pattern described by Hoey (1983, 2001) and suggests corpus-based approaches for helping students perform the functions of proposing and evaluating a solution. Van Rij-Heyligers (2011) investigates how RAs in the sociology of education carry out the function of indicating a gap in research. She stresses the diversity of realisations possible and argues that corpus-based work has the potential to bring a "critical genre dimension" into the analysis and teaching of genres (2011, 133). In earlier work (Charles 2007, 2011a), I described a course for graduate writers, which was designed to teach discourse functions, including, for example, making claims, indicating a gap in research and defending research from criticism. The materials employed a top-down approach: individual discourse functions were first presented and studied in the form of paper-based texts and this stage was followed by hands-on concordancing using a small corpus of theses (approximately 500,000 words). Corpora offer an appropriate means of investigation because discourse functions can be associated with typical linguistic realisations and these elements can be used as probes to locate and examine the function under study. Work on discourse issues, then, offers another way to counter the overly lexico-grammatical focus of much corpus-based pedagogy.

In this chapter I present and discuss four pedagogical tasks designed to show how corpus-based methods can be used to focus on genre and discourse issues. The first two tasks focus on the genre of the thesis, illustrating how paper-based materials using concordance data can be used to help students take a critical approach to one move of the introduction. This task makes use of Bunton's (2002) model of thesis introductions, as described in Secion 3.2 below. The remaining two tasks focus on discourse functions and show how students can make use of a self-compiled personal corpus to examine the functions of Indicating a Gap in Research and Defending your Research from Criticism. These two functions may be difficult for students to perform, but are crucial for successfully positioning research as a worthwhile contribution to the field.

## 2. Pedagogical context

The four tasks presented below share the same overall pedagogical context, although they took place on two separate courses. The study described in this chapter was carried out within the programme on academic writing for graduate students at Oxford University Language Centre. The programme is organised into several parallel multidisciplinary classes and caters for students from many different nationalities and language backgrounds. Students are at an advanced level, with an average score of 7 on IELTS. Courses last for 6-8 weeks and consist of one 2-hour session per week, with an attendance of around 16 participants per class. A questionnaire on students' backgrounds and needs administered in the academic year 2010 – 2011 was completed by 140 participants. It showed a roughly equal split between doctoral students (54%) and Master's students (46%). The two genres that the students most needed to write were reported as the thesis/dissertation (96%) and the research article (71%).

## 3. Investigating genre: Introductions in theses

Tasks 1 and 2 form part of a unit on introductions from the course on *Writing a Thesis or Dissertation*. This course is based on research descriptions of some of the part-genres of the thesis/dissertation, including, for example, the introduction, conclusion and abstract. The teaching material presents students with extended examples and data taken from a corpus of theses in two contrasting disciplines: 10 doctoral theses in materials science, a hard natural science (396,000 words) and 10 MPhil theses[1] in politics, a soft social science (255,000 words). In total, then, the corpus amounts to over half a million words.

### 3.1. General pedagogical procedure

The pedagogical procedure for all units is summarised below:

*Phase 1. Awareness-raising: Reflection and Expression*
The first step is for students to reflect on and note down the features they expect to characterise the part-genre. They discuss their ideas in pairs or small groups and this is followed by feedback and discussion with the whole class. Students' suggestions are listed for the whole class to see.

*Phase 2. Presentation*
In the second phase, the model found by research on genre analysis is presented and compared with the students' suggestions. The whole class discusses and evaluates the model.

*Phase 3. Genre Analysis*
The model is applied by the students to an extended example from the corpus. Students discuss their analyses in pairs/small groups; whole class feedback and discussion then take place.

*Phase 4. Extension*
In this phase a variety of tasks are assigned, depending on the difficulties of the part-genre and the needs of the class. For example these tasks may highlight variation in the part-genre and/or compare and contrast the studied part-genre with other part-genres. Detailed analysis of specific moves and steps and/or the study of lexico-grammar may also be undertaken.

*Phase 5. Application*
For homework, students write an example of the part-genre in their own field.

## 3.2. Extension task 1: Variation in step 1 of the thesis introduction

The two tasks I will discuss here are examples of the Extension phase, which is particularly suited to corpus-based work. The first highlights variation within the part-genre, while the second compares and contrasts the introduction to the whole thesis with introductions to individual chapters. However, before examining these tasks in detail, it is necessary to introduce the model upon which they are based. The model presented in Phase 2 of the material on introductions is proposed by Bunton (2002) and builds on Swales' (1990) well-known Create a Research Space model for RA introductions. Drawing on an analysis of the introductions to 45 PhD theses in a wide range of disciplines, Bunton presents the most frequently occurring moves and steps of the part-genre as follows:

*Move 1: Establishing a Territory*
STEPS
1: Claiming centrality
2: Making topic generalisations and giving background information
3: Defining terms (Engineering, Arts, Social Sciences)
4: Reviewing previous research

*Move 2: Establishing a Niche*
STEPS
1A: Indicating a gap in research
1B: Indicating a problem or need
1C: Question-raising (Social Sciences, Arts)
1D: Continuing a tradition (Medicine, Social Sciences)

*Move 3: Announcing the Present Research (Occupying the Niche)*
STEPS
1: Purposes, aims or objectives
2: Work carried out (Science, Engineering)
3: Method
4: Materials or Subjects
5: Findings or Results
6: Product of research (Engineering)/Model proposed (Social Sciences)
7: Significance/Justification
8: Thesis structure

(Bunton 2002, 74)

While Bunton's model provides an excellent starting point for teaching the introduction, if students are to become more aware of the diversity of introduction types, they need to be exposed to multiple attested examples. Moreover, in Phase 1 of the pedagogical procedure outlined above, students often proposed steps that did not occur in the model and were concerned about whether their suggestions were valid and could be used. One of the aims of Task 1, then, was to hold the model up to scrutiny by showing the students a wider range of possibilities. Although corpus data is very well-adapted to the task of providing multiple examples, there is a practical problem: whole introductions can extend over several pages and thousands of words. Where the aim is to highlight variation in a part-genre, this potentially confronts the students with a large amount of reading, which is time-consuming and unwieldy for classroom use.

It is necessary, then, to find a way of maximising the number of examples presented, while minimising the amount of text to be read. One option for dealing with this problem is to use a single Move to show the variation and to present only a limited amount of text for each example. As students are often concerned about how to begin the thesis, it was decided to focus on Bunton's Move 1 and in order to cut down the amount to be read, to present just the initial sentence of each thesis. It was hypothesised that the first sentence might reveal substantial information about how the writer chooses to situate the thesis within the field and that

these sentences could form the basis of a task which would indicate to students the extent of conformity to and diversity from Bunton's model.

Table 1 presents the analysis of the steps performed in the first sentence of the 20 introductions.

| Step Type | Move and Step in Bunton's Model | Number of Instances in Sentence 1 |
|---|---|---|
| Claiming centrality | Move 1 Step 1 | 4 |
| Making topic generalisations and giving background information | Move 1 Step 2 | 13 |
| Referring to previous research[2] | Move 1 Step 4 | 3 |
| Indicating a problem or need | Move 2 Step 1B | 1 |
| Question-raising | Move 2 Step 1C | 2 |
| Purposes, aims or objectives | Move 3 Step 1 | 2 |
| Work carried out | Move 3 Step 2 | 1 |
| Anecdote | Not present | 1 |
| Description of world event | Not present | 2 |
| **Total** | | **29** |

**Table 1. Analysis of steps performed in initial sentences of 20 theses.**

It is clear that Bunton's Move 1 Steps 1 and 2 (Topic Generalisation and Centrality Claim) along with Step 4 (Reviewing/Referring to Previous Research) feature highly in these initial sentences, accounting for two-thirds of the steps identified. However two new steps are distinguished, which do not figure in Bunton's model: Anecdote and Description of World Event. These are similar to each other, in that both recount what is essentially a story and are probably chosen because they encapsulate an important message of the thesis, presenting it in an immediately accessible and appealing way. The steps differ, however, in that an Anecdote tends to have a personal resonance with the writer of the thesis, while the Description of a World Event uses an international occurrence as the basis of the step. Clearly the latter option is limited to those disciplines in which world events play a crucial role. These two options are exemplified below:

**(1)**
**Anecdote**
**One of my favourite possessions as a boy was** a crystal of copper sulphate which **I had grown**. (Materials Science)

**(2)**
**Description of World Event**
**On 27 July 1950,** the Canadian cabinet approved the participation of Canadian troops in the Korean War. (Politics)

Perhaps more importantly, the analysis of these opening sentences revealed a complexity which proved insightful for pedagogical purposes. First, 50% of the sentences performed more than one step, which highlighted the fact that steps can be extremely short in extent. Moreover, a third of the total of the 29 steps identified do not appear as part of Bunton's Move 1. Question-raising and Indicating a Problem or Need appear in Move 2 and Purposes, Aims or Objectives and Work Carried Out in Move 3. What this suggests is that there is a much wider range of possible ways of beginning a thesis than those offered by the model. This variation and complexity are illustrated in the examples below:

**(3)**
**Aim + Question-raising**
This thesis **aims** to make a first cut at answering the following question: **why** do states come to resemble one another as they interact in the international system? (Politics)

**(4)**
**Topic Generalisation + Centrality Claim + Problem + Reference to Previous Research**
The Supreme Court is an institution **central** to the American polity, **yet** as **Bailey points out**, "the way the Court is studied stands **outside the mainstream of political science**"[1]. (Politics)

**(5)**
**Work Carried Out**
**This thesis describes the work carried out** to investigate the high temperature erosion resistance of uncoated and coated thermal barrier tiles. (Materials Science)

In Task 1, then, students are given a selection of these introductory sentences and asked to identify the steps and to create new labels for any steps that are not covered by the model. They are requested to notice not only the move to which the steps belong, but also the way in which they can be combined even within relatively short stretches of text. This naturally leads on to an evaluation of ways to begin a thesis, offering the opportunity for students to articulate their ideas and the reasons that underlie their preferences. In a multidisciplinary class this provides useful material for discussion and comparison of disciplinary tendencies.

Working with multiple examples from introductions enhanced students' awareness of reader needs and expectations, and the differing ways in which these can be satisfied. From discussions with peers, learners gained a better understanding of how disciplinary norms affect these writer choices.

### 3.3. Extension task 2: Contrasting thesis and chapter introductions

Task 2 provides follow-up to Task 1 and attempts to expand genre descriptions of the introduction further. It aims to compare and contrast the introduction to the whole thesis with the introductions to individual chapters. In research on the thesis/dissertation, it is noticeable that little attention has been paid to the hierarchical organisation of the genre. While individual part-genres like the introduction (Bunton 2002), the conclusion (Bunton 2005) and the literature review (Kwan 2006; Thompson 2009) have been analysed, little research has been carried out to determine how individual chapters in the body of the thesis are internally structured and linked together. There may be an assumption that such chapters conform to genre descriptions of the RA. However, this is to ignore the fact that individual thesis chapters are part of a much longer document and thus each chapter has to be situated not just within the discipline, but within the text as a whole. Currently there seems to be little evidence about how this is done, although Bunton (1999) draws attention to the importance of higher level metatext as a key feature of theses.

Investigation of the introductions to the individual chapters in the corpus reveals that the Topic Generalisation is the most frequent step used. However two other steps also occur that are not present in Bunton's (2002) model: Statement of the Aim or Content of the Present Chapter and Referring Back to Previous Chapters. Concordance lines illustrating these two functions in the corpus are given below.

**Aim or Content of the Present Chapter**

1. This chapter aims to accomplish three objectives. Firstly it will introduce the reader to
2. This chapter aims to give a general introduction to grain growth and the relevance of this
3. The aim of these experiments is to explore the effect of solute on triggering of secondary
4. This chapter describes the experimental techniques and apparatus used to investigate the
5. This chapter describes an investigation of secondary recrystallisation in the presence of
6. This chapter describes work carried out to investigate the ability of the 3D Monte Carlo P
7. This section describes an attempt to model Zener pinning at f<0.01.We model a boundary m

8. This chapter will investigate and compare a wide range of semiconductor material using SE-

9. The results presented in this chapter relate to the parameters affecting the incident beam

10. In this chapter, the methods used by Segal and Cover (1989) are briefly outlined. A simi

11. The task of this chapter is to determine the extent to which British defence policy was i

### Referring Back to Previous Chapters

1. As discussed in Chapter II, it was suggested that the Gas Reaction Cell microscope could b

2. As demonstrated in the previous chapter, the two fundamental rules of international law we

3. As suggested in Chapter II, the Gas Reaction Cell could be used to investigate the in-situ

4. In Chapter IV, graphite was investigated in the Gas Reaction Cell, and its remarkable beha

5. The erosion conditions expected within a gas turbine combustor were described in Section 2.

6. The previous chapter provides evidence that, in significant cases, an 'attitude' based mod

7. The last few chapters of this thesis have been concerned primarily with carbon blacks, and

Such concordance lines provide useful material for pedagogic tasks. In Task 2 the two sets of lines are jumbled and students are asked to sort them by identifying two steps and allocating each line appropriately. Further questions require the students to contrast these chapter introductions with the thesis introductions in Task 1, to discuss the differences between the two types of introduction and to suggest reasons for what they have observed. Finally students focus on the lexico-grammar of the two steps of chapter introductions. They are asked to write down two different ways of stating an aim, describing the content and referring back to previous chapters and to notice the most frequent tense chosen to perform each step. In pairs, they compare their findings and explain the tense choices they see. These lexico-grammatical tasks are designed to link the generic steps to their linguistic realisations, thereby helping students with the Application Phase, in which they write the first paragraph of an introduction. In differing ways, then, all the tasks described in this section enlarge upon the options available to students for constructing introductions. Rather than presenting genre descriptions as fixed and rigid models to be strictly adhered to, they seek to problematise genre descriptions, expand them and encourage students to become aware of a wider set of possibilities. In carrying out such tasks, students can verify and modify existing descriptions, which leads them to understand better the options available and thus enables a more informed choice in their own writing.

## 4. Investigating discourse functions: Research articles

The previous two tasks use paper-based materials and rely on the availability of pre-constructed corpora and prior analysis by the teacher. The two that follow take a hands-on approach, in which the responsibility for building and investigating the corpora are assumed by the learners.

This work is carried out as part of the academic writing programme described above in Section 2. The aim is to teach participants how to construct their own personal corpus of RAs and to investigate it by focusing on discourse functions in their own discipline. Corpus searches were used to identify such functions and to link them to specific lexico-grammatical choices. In their account of using do-it-yourself (DIY) corpora, Lee and Swales (2006, 57) suggested that the needs of advanced-level graduate students were primarily for "lexico-grammatical fine-tuning". While accepting that attention to lexico-grammatical detail is certainly necessary, I would concur with Starfield (2004) that such students also require work on text structure. The focus on discourse functions, then, was designed to show students how relatively long stretches of text may be organised in fairly conventional ways.

The 6-week course uses the AntConc software (Anthony 2011) combined with searches assigned by the teacher and sets of focusing questions to help the students notice the characteristics of given discourse functions in their corpus. Full details of the procedure used for examining discourse functions are available in Charles (2007, 2011a), while a description and evaluation of this DIY corpus building course can be found in Charles (2012). Here I will focus on two examples which show the potential of personal DIY corpora to provide students with discipline-specific information on how discourse functions are performed in their field.

In 2011, 47 students took the course and built personal corpora from RAs which they had already downloaded for their own research purposes. Thus one of the main criteria for inclusion was relevance to the student's own research topic. In consultation with their supervisors, students were also advised to select RAs considered to be well-written and well-regarded in the field and to include papers from a range of different writers. The corpora ranged in size from 6 to 164 files, with an average of 17 files per corpus, where each file was a single RA. Not all students reported the number of words in their corpus, but 36 students constructed corpora which ranged in size from 10,000 to well over a million words; the largest was 1,631,564 words. The average number of words per corpus was 100,406. Although most of the corpora were small by research standards,

the advantages of using such corpora in EAP teaching have been noted by several authors (e.g. Aston 2002; Tribble 2002; Flowerdew 2004). For example, they can be compiled to address the needs of specific groups of students in terms of genre and discipline and, as learners are familiar with the context of the corpus texts, they can more readily understand and interpret the concordance lines, while not being overwhelmed by a large quantity of data. In multi-disciplinary classes, the ability to consult discipline-specific corpora is of particular importance.

### 4.1. Investigating the discourse function of "Indicating a Gap in Research"

The first function to be addressed here is Indicating a Gap in Research, a variant of the Gap in Knowledge-Filling Pattern (Hoey 2001) characteristic of academic discourse. This function occurs in several different genres and part-genres, most notably in both the thesis and the research article, where it is attested in introductions (Swales 1990, 2004; Bunton 2002) and literature reviews (Kwan 2006; Flowerdew and Forest 2009). Bunton (2005) also occasionally found what he calls the "Gap/niche" step in conclusions to theses as part of the "Introductory Restatement" Move. Following the work of Flowerdew and Forest (2009), who identified the keywords of PhD literature reviews in applied linguistics, students were asked to search on the terms: *study, studies,* and *research, work* and *literature*. They were then instructed to use the Plot Tool to locate potential instances of the function.

The Plot Tool provides a graphic representation of all the instances of a search term in a corpus, with each file shown as a horizontal bar and vertical lines to indicate each individual instance. Because the RA is a relatively conventionalised genre, this view allows the user to get a rough idea of the position of the search term in the generic structure. Thus instances of the search term towards the beginning of a file are likely to be in the abstract or introduction, while those towards the end will probably be in the conclusion. In this way, likely sites of interest can be identified even in a corpus that is not tagged according to its generic sections. Having identified potentially relevant instances, students were asked to access the original file from the Plot Tool, to examine the text and to ascertain whether the search term functioned to indicate a gap in research. Each student worked on their own personal corpus individually and then discussed their results with fellow-students.

Working with a corpus of 221,636 words (20 RA files) in anthropology, Anne[3] noticed that a Plot of the term *research* showed

several occurrences towards the ends of files. This was rather unexpected, as it was thought that *research* would be more likely to occur towards the beginning of files as part of the Establishing a Niche Move in the Introduction or Literature Review. On further study of the original files, Anne found the following examples of this noun in the conclusions to the RAs:

**(6)**
The project, according to Hsii, its author, **awaits further research**.

**(7)**
**It remains open for further research** how successful these practices are.

**(8)**
**The challenge for future research** is to document the different motivations.

**(9)**
[…] **future research in these areas** has the potential to contribute significantly to reassessing the interpretive tradition.

**(10)**
Taken together, the evidence from these different cases raises several questions pointing to **the need for more empirical research** to assess whether standardization […]

Here, the noun does indeed form part of the function of Indicating a Gap in Research, but, in contrast to the findings of Bunton (2005), it is used to point to the necessity for further work, rather than as a restatement of the rationale for the study. The gap thus remains open, and may be elaborated upon with information about potential benefits to be gained or tasks to be carried out.

Using her own corpus, then, Anne was able to identify this discourse function and to see how it is used in anthropology RA conclusions for suggesting future work. Moreover, the examples she found are rich in lexico-grammatical chunks that are useful for performing the function in her own writing (e.g. *X awaits further research; The challenge for future research*).

## 4.2. Investigating the discourse function of "Defending your Research against Criticism"

In examining the function of Defending your Research against Criticism, corpus tasks focused on the use of concessions. In a study of the British Academic Written English corpus (BAWE), Charles (2011b) examined student genres in business studies, computer science, chemistry and politics. She investigated the use of contrast and concession subordinators (*although, though, while, whilst* and *whereas*) and found that sentence initial *although/though* and *while/whilst* are often used to signal concessions. These subordinators form part of a semantic sequence (Hunston 2008) in which the subordinate clause evaluates positively and the main clause negatively; a reason for the negative evaluation may also be given:

***although/though/while/whilst*** + **POSITIVE EVALUATION** + **NEGATIVE EVALUATION** + **REASON** **(optional)**

**(11)**
**Although the first three tests were successful**, the **fourth was not**. **The reason for this was** that when using the strictly greater than sign, neither of the threes is the maximum. (6101e: Computer Science)

Charles suggested that one reason for the occurrence of this sequence is that student genres require their writers to balance critical and positive comments about their own work. They need to show their awareness of its possible flaws, but at the same time, must guard against undermining its value completely. In these concessions, writers first note an achievement of their work, then anticipate and forestall the reader's potential criticism by mentioning a shortcoming and optionally offering a reason to explain its presence. In this way, the concession functions to strengthen the writer's point and to defend their work from criticism.

The aim of the task on this function was to get students to investigate concessions in their corpora of RAs and to ascertain whether concessions performed a similar function in these expert texts. Students were asked to retrieve concordance lines for sentence initial *although* and *while*, to access the original files and to determine whether the examples were concessions. If so, they were asked to check whether the examples followed the positive – negative sequence and what function they were performing in the text. Working with a corpus of 90,708 words in zoology (16 RA files), Jan found 49 hits for sentence initial *although*. These often occurred in the Results and Discussion section, but although many marked

concessions, he found roughly equal numbers of two evaluative sequences: positive - negative and negative - positive. Examples are given below:

**(12)**
**Positive - Negative**
**Although** we found that DNA barcodes are likely to be **very useful** in identifying a newly encountered aphid specimen to species […], we note that **identification cannot reliably be extended** to deeper levels […]

**(13)**
**Negative - Positive**
**Although** some aphids in this study **are infected** with additional, facultative symbionts that **can affect tolerance to heat** [6,14], **our comparisons controlled for differences** in facultative symbiont infections.

Although the function of both sequences is to defend the writer's work from criticism, they operate slightly differently from the examples in the BAWE student texts. Using his knowledge of the discipline, Jan explained positive - negative sequences like example (12) by noting that in zoology it is necessary to establish the boundaries of a claim very carefully; he accounted for negative - positive sequences like example (13) by pointing to the importance of demonstrating the reliability of the data. Thus Jan was able to verify the use of concessions in his discipline and to gain an insight into the way they can be used in two different ways defend and thereby strengthen the writer's work.

These tasks show that personal DIY corpora have the potential to make the discourse choices of expert writers accessible and visible to students, who can then make use of this knowledge in their own writing. In the spirit of Johns (1991b) the students acted as researchers in their own discipline and were able to gain insights from their corpora which were directly relevant to their own needs. Thus personal corpora allow for individualisation within a classroom setting, and, of equal importance, help students to become more independent in their learning.

## 5. Conclusion

In this study I have attempted to show how corpora can be used to enrich EAP pedagogy by facilitating the study of genre and discourse issues in academic writing. I have illustrated two approaches. The first uses traditional paper-based materials derived from prior analysis of a

corpus and shows how such tasks can contribute to raising student awareness of the variability of genres. Although some degree of corpus familiarity is needed on the part of the teacher, readily available corpora such as BAWE or MICUSP (Michigan Corpus of Upper-level Student Papers) can provide an extremely useful resource for constructing this type of pedagogic material. The second type of approach, in which students compile their own personal corpora, requires both more extensive computer facilities and greater corpus expertise on the part of the teacher. However, where these resources are available, such an approach provides a viable way of addressing the variability of academic writing across different disciplines and of fostering student independence in the learning process. Such personal corpora also have the advantage of providing ongoing, long term support for student writers (Charles 2013).

Over the past decades, corpus-based research has led to enormous gains in our understanding of academic written discourse, but attention to corpus-based pedagogy has somewhat lagged behind. This study has only scratched the surface in terms of the possible applications of corpora to the teaching of genre and discourse. Further research efforts are needed in order to achieve richer outcomes for students in terms of participation, understanding and autonomy.

## Notes

[1] The MPhil degree requires two years of study and is assessed by a thesis which should be an original piece of research totalling approximately 30,000 words.
[2] As this analysis only examines the initial sentence of each introduction, I have modified Bunton's step of Reviewing Previous Work to become Referring to Previous Research. While it would not be possible to review research within a single sentence, a reference to previous research was considered to be an important element and one that overlaps with Bunton's category.
[3] Names of students are pseudonyms.

## References

Anthony, Laurence. 2011. *AntConc* (Version 3.2.4). Tokyo, Japan, Waseda University. http://www.antlab.sci.waseda.ac.jp/.

Aston, Guy. 1995. Corpora in language pedagogy: Matching theory and practice. In *Principle and practice in applied linguistics*, ed. Guy Cook and Barbara Seidlhofer, 257-270. Oxford: Oxford University Press.

—. 2002. The learner as corpus designer. In *Teaching and learning by doing corpus analysis*, ed. Bernhard Kettemann and Georg Marko, 9-25. Amsterdam: Rodopi.

Bernardini, Silvia. 2000. Systematising serendipity: Proposals for concordancing large corpora with language learners. In *Rethinking language pedagogy from a corpus perspective*, ed. Lou Burnard and Tony McEnery, 225-234. Frankfurt: Peter Lang.

—. 2002. Exploring new directions for discovery learning. In *Teaching and learning by doing corpus analysis*, ed. Bernhard Kettemann and Georg Marko, 165-182. Amsterdam: Rodopi.

Bianchi, Francesca, and Roberto Pazzaglia. 2007. Student writing of research articles in a foreign language: Metacognition and corpora. In *Corpus linguistics 25 years on*, ed. Roberta Facchinetti, 261-287. Amsterdam: Rodopi.

Bondi, Marina. 2001. Small corpora and language variation. In *Small corpus studies and ELT*, ed. Mohsen Ghadessy, Alex Henry and Robert Roseberry, 135-174. Amsterdam: John Benjamins.

Boulton, Alex. 2010. Data-driven learning: Taking the computer out of the equation. *Language Learning* 60: 534-572.

British Academic Written English Corpus (BAWE). http://wwwm.coventry.ac.uk/researchnet/BAWE/Pages/BAWE.aspx

Bunton, David. 1999. The use of higher level metatext in PhD theses. *English for Specific Purposes* 18(S): S41-S56.

—. 2002. Generic moves in PhD thesis introductions. In *Academic discourse*, ed. John Flowerdew, 57-75. London: Longman.

—. 2005. The structure of PhD conclusion chapters. *Journal of English for Academic Purposes* 4: 207-224.

Charles, Maggie. 2007. Reconciling top-down and bottom-up approaches to graduate writing: Using a corpus to teach rhetorical functions. *Journal of English for Academic Purposes*6: 289-302.

—. 2011a. Using hands-on concordancing to teach rhetorical functions: Evaluation and implications for EAP writing classes. In *New trends in corpora and language learning*, ed. Ana Frankenberg-Garcia, Lynne Flowerdew and Guy Aston, 26-43. London: Continuum.

—. 2011b. Making concessions in academic writing: A corpus study of patterns and semantic sequences. In *Proceedings of the Corpus Linguistics Conference 2011*, ed. Nicholas Groom and Oliver Mason. Birmingham, UK: University of Birmingham. http://www.birmingham.ac.uk/documents/college-artslaw/corpus/conference-archives/2011/Paper-88.pdf

—. 2012. "Proper vocabulary and juicy collocations": EAP students evaluate do-it-yourself corpus-building. *English for Specific Purposes* 31: 93-102.

—. 2013. Student corpus use: Giving up or keeping on? In *TaLC10: Proceedings of the 10th International Conference on Teaching and Language Corpora*, ed. Agnieszka Lenko-Szymanska. Warsaw: Warsaw University Press. Available online at: http://talc10.ils.uw.edu.pl/proceedings/.

Cortes, Viviana. 2007. Genre and corpora in the English for academic writing class. *ORTESOL Journal* 25: 9-16.

—. 2011. Genre in the academic writing class: With or without corpora? *Quaderns de Filologica. Estudis Lingüistics* XVI: 65-79.

Cresswell, Andy. 2007. Getting to "know" connectors? Evaluating data-driven learning in a writing skills course. In *Corpora in the foreign language classroom*, ed. Encarnacion Hidalgo, Luis Quereda and Juan Santana, 267-287. Amsterdam: Rodopi.

Estling Vannestål, Maria, and Hans Lindquist. 2007. Learning English grammar with a corpus: Experimenting with concordancing in a university grammar course. *ReCALL* 19: 329-350.

Flowerdew, John, and Richard Forest. 2009. Schematic structure and lexico-grammatical realization in corpus-based genre analysis: The case of research in the PhD literature review. In *Academic writing: At the interface of corpus and discourse*, ed. Maggie Charles, Diane Pecorari and Susan Hunston, 15-36. London: Continuum.

Flowerdew, Lynne. 1998. Corpus linguistic techniques applied to textlinguistics. *System* 26: 541-552.

—. 2004. The argument for using English specialized corpora to understand academic and professional language. In *Discourse in the professions: Perspectives from corpus linguistics*, ed. Ulla Connor and Thomas Upton, 11-33. Amsterdam: John Benjamins.

—. 2005. An integration of corpus-based and genre-based approaches to text analysis in EAP/ESP: Countering criticisms against corpus-based methodologies. *English for Specific Purposes* 24: 321-332.

—. 2008. *Corpus-based analyses of the problem-solution pattern*. Amsterdam: John Benjamins.

—. 2009. Applying corpus linguistics to pedagogy. *International Journal of Corpus Linguistics* 14: 393-417.

—. 2012. Exploiting a corpus of business letters from a phraseological, functional perspective. *ReCALL* 24: 152-168.

Frankenberg-Garcia, Ana. 2005. A peek into what today's language learners as researchers actually do. *International Journal of Lexicography* 18: 335-355.

Frankenberg-Garcia, Ana, Lynne Flowerdew, and Guy Aston. eds. 2011. *New trends in corpora and language learning*. London: Continuum.

Freedman, Aviva. 1993. Show and tell? The role of explicit teaching in the learning of new genres. *Research in the Teaching of English* 27: 222-251.

Gavioli, Laura. 2005. *Exploring corpora for ESP learning*. Amsterdam: John Benjamins.

Granath, Solveig. 2009. Who benefits from learning how to use corpora? In *Corpora and language teaching*, ed. Karin Aijmer, 47-65. Amsterdam: John Benjamins.

Hafner, Christoph, and Christopher Candlin. 2007. Corpus tools as an affordance to learning in professional legal education. *Journal of English for Academic Purposes* 6: 303-318.

Hoey, Michael. 1983. *On the surface of discourse*. London: George Allen and Unwin.

—. 2001. *Textual interaction*. London: Routledge.

Hunston, Susan. 2008. Starting with the small words: Patterns, lexis and semantic sequences. *International Journal of Corpus Linguistics* 13: 271-295.

Johns, Tim. 1991a. From printout to handout: Grammar and vocabulary teaching in the context of data-driven learning. In *Classroom concordancing*, ed. Tim Johns and Philip King, 27-37. Birmingham: ELR University of Birmingham.

—. 1991b. Should you be persuaded: Two samples of data-driven learning materials. In *Classroom concordancing*, ed. Tim Johns and Philip King, 1-16. Birmingham: ELR University of Birmingham.

Kaltenböck, Gunther, and Barbara Mehlmauer-Larcher. 2005. Computer corpora and the language classroom: On the potential and limitations of computer corpora in language teaching. *ReCALL* 17: 65-84.

Kübler, Natalie. ed. 2011. *Corpora, language, teaching, and resources: From theory to practice*. Bern: Peter Lang.

Kwan, Becky. 2006. The schematic structure of literature reviews in doctoral theses of applied linguistics. *English for Specific Purposes* 25: 30-55.

Lee, David, and John Swales. 2006. A corpus-based EAP course for NNS doctoral students: Moving from available specialized corpora to self-compiled corpora. *English for Specific Purposes* 25: 56-75.

Michigan Corpus of Upper-level Student Papers (MICUSP). 2009. Ann Arbor, MI: The Regents of the University of Michigan. http://micusp.elicorpora.info/

Mudraya, Olga. 2006. Engineering English: A lexical frequency instructional model. *English for Specific Purposes* 25: 235-256.

Starfield, Sue. 2004. "Why does this feel empowering?" Thesis writing, concordancing and the corporatizing university. In *Critical pedagogies and language learning*, ed. Bonny Norton and Kelleen Toohey, 138-157. Cambridge: Cambridge University Press.

Swales, John. 1990. *Genre analysis*. Cambridge: Cambridge University Press.

—. 2002. Integrated and fragmented worlds: EAP materials and corpus linguistics. In *Academic discourse*, ed. John Flowerdew, 150-164. London: Longman.

—. 2004. *Research genres*. Cambridge: Cambridge University Press.

Thomas, James, and Alex Boulton. eds. 2012. *Input, process and product: Developments in teaching and language corpora.* Brno: Masaryk University Press.

Thompson, Paul. 2009. Literature reviews in applied PhD theses: Evidence and problems. In *Academic evaluation*: *Review genres in university settings*, ed. Ken Hyland and Giuliana Diani, 50-67. Basingstoke: Palgrave Macmillan.

Thurstun, Jennifer, and Christopher Candlin. 1997. *Exploring academic English: A workbook for student essay writing*. Sydney: NCELTR.

Tribble, Christopher. 2002. Corpora and corpus analysis: New windows on academic writing. In *Academic discourse*, ed. John Flowerdew, 131-149. London: Longman.

van Rij-Heyligers, Josta. 2011. Breaking the chains of rhetorics in academia: Corpus-based research as tool for transformation in discourse? In *Corpora, language, teaching, and resources: From theory to practice*, ed. Natalie Kübler, 133-152. Bern: Peter Lang.

Weber, Jean-Jacques. 2001. A concordance- and genre-informed approach to ESP essay writing. *ELT Journal* 55: 14-20.

Yoon, Choongil. 2011. Concordancing in L2 writing class: An overview of research and issues. *Journal of English for Academic Purposes* 10: 130-139.

Yoon, Hyunsook. 2008. More than a linguistic reference: The influence of corpus technology on L2 academic writing. *Language Learning and Technology* 12: 31-48.

# CHAPTER TWELVE

# TEXT AND CORPUS: MIXING PARADIGMS IN EAP SYLLABUS AND COURSE DESIGN

MARIA FREDDI
UNIVERSITÀ DI PAVIA, ITALY

## 1. Introduction

The particular demands EAP teaching poses for teachers in the higher education system call for continuous reflection on course design. Global tendencies such as growing students mobility and the fact that English is increasingly being chosen as the medium of instruction by many European universities across degree courses are at odds with local scenarios where university students arrive from school lacking language proficiency in English and little space is devoted to foreign language education in academic curricula. For these reasons alone, EAP instruction offers an indispensable form of students empowerment in its multiple objectives and broad educational scope, covering internationalization, the promotion towards membership of an academic community, the development of study skills in what is the lingua franca of professional communication, and consequently opening up occupational possibilities for those instructed in this way. This is why a growing body of research is concerned with descriptions of academic practices across disciplinary fields (e.g. Hyland 2012a; Nesi and Gardner 2012), the modes of discourse or genres in which specialised knowledge is communicated and the linguistic features characteristic of these genres (e.g. Hyland 2000; Hyland and Bondi 2006; Bondi 2008), while pedagogies of improved relevance for new cohorts of students as well as language teachers are being developed all over the world. Such pedagogies can no longer rest on the tacit acquisition of unspoken skills, what has been dubbed a "pedagogy of

osmosis" (Turner 2011, 21 cited in Nesi and Gardner 2012, 261). Rather, it is subject to experimentation and explicit evaluation.

Various pedagogical practices have been reported in the literature. Some stem directly from genre studies and the tradition fruitfully established by Swales (1990) and subsequent work (e.g. Johns 1997; Peck McDonald 2004; Bruce 2011), others have experimented with corpus methods especially in the fashion of "data-driven learning" (e.g. Bernardini 2004, and recently through the web as corpus approach Gatto 2014). While the former are focused on familiarizing students with various genres of communication, their rhetorical staging and linguistic realization, usually taking what identifies as a top-down approach to text, i.e. from function to form (cf. Flowerdew 2012) and are often geared towards developing writing skills (e.g. Feak and Swales 2011; Swales and Feak [1994] 2012), the latter exploit the increased possibilities brought about by electronic corpora both as a series of theoretical statements and as an approach to the study of language, including domain-specific language (Coxhead 2000; Cobb and Horst 2001; Hyland 2006, 2012b; Thompson 2006, 2007; Krishnamurty and Kosem 2007). In the main, these have made use of concordancing as a tool for academic and specialised language learning. This approach is usually bottom-up, i.e. from form to function, as it promotes noticing and discovery of forms of academic language repeated regularly across a body of texts (cf. Flowerdew 2002, 2004, 2009; Bondi and Diani 2009). Because of the need to address both functional and formal concerns, attempts have been made to integrate a genre teaching methodology with a corpus teaching methodology (e.g. Charles 2007, 2011, 2012; Flowerdew 2009; Diani 2012) and in this context the advantages and disadvantages of bringing the concordance lines to the EAP classroom (e.g. time management, accessibility, helpfulness and learning outcomes) have been evaluated (notably by Charles 2012, 2014).

Corpora have proven relevant to EAP instruction also in a more indirect way as corpus findings have been brought to bear on course and materials design. However, after an initial surge of proposals following the diffusion of corpus data in the early nineties (cf. Flowerdew 1993; Thurstun and Candlin 1998), corpus-based courses appear to have stalled and corpus-informed syllabi have only recently gained new popularity, which is reflected in a number of pedagogic wordlists (e.g. the Academic Word List and Academic Phrase List drawn by Coxhead 2000, 2010) and in some new EAP coursebooks published simultaneously by Cambridge University Press and Oxford University Press in 2012 (see the *Cambridge Academic English* series and the *Oxford EAP* series). Both coursebooks

are based on corpus data and choose and organise their contents in light of corpus findings. Similarly, Swales and Feak ([1994] 2012) in the latest edition of their widely diffused coursebook *Academic Writing for Graduate Students: A Course for Non-Native Speakers of English* have made use of data and information extracted from the MICUSP, the Michigan Corpus of Upper-Level Student Papers containing advanced students' writings, to design tasks and materials to be used in the writing class.

All these corpus-informed syllabi and materials, however, address what is understood as English for General Academic Purposes (EGAP) rather than targeting skills and language which are related to the demands of a particular discipline or Department (English for Specific Academic Purposes, ESAP). Indeed, there seems to be little material available specific to the Humanities where both genre theory and corpus data are used to inform the development of a syllabus, interpret the communicative practices involved in this particular context, describe the disciplinary textuality (see section 5 below), and choose and use the target texts.

Within this perspective, the present study makes the case for a mixed text-corpus approach to academic English to develop a syllabus in EAP reading for the Humanities. The argument is made for referring to corpus methods to guide the instructors and course developers' choices without necessarily bringing corpus observation tools into class (unlike the approach taken, among others, by Charles 2007, 2011). The case considered is a course taught by the author for the past five years at her home University aimed at undergraduate Humanities students and titled *Reading Skills in English for the Humanities*. However, the specific case study is discussed in the more general terms of quality control as the planning of a process (ReVelle et al. 1998), with the aim of providing a wider analytical model that applies to other settings. Quality control is hereinafter used to conceptualise each stage in the design of the course from preparing to implementing and to evaluating it. Through analysis of the whole design process, including students' response data, the efficacy of some of the choices made in designing the syllabus are questioned and suggestions for further amendments are put forward.

## 2. Syllabus design

A syllabus is understood as "the specification of aims and the selection and grading of content to be used as a basis for planning foreign language, or any other educational, courses" (Newby 2004, 590). With special reference to (foreign) language courses, Jordan ([1997] 2012) explains:

> Designing a syllabus involves examining needs analysis and establishing goals. It then entails the selection, grading and sequencing of the language and other content, and the division of the content into units of manageable material. (Jordan [1997] 2012, 56)

In order to implement the syllabus, a series of methodological choices are to be made as to the selection and/or development of materials, tasks, activities and exercise types, thus the syllabus can be seen as a means of influencing materials design and classroom practices.

There are usually several stages to the design of a syllabus, specifically its (i) preparation, (ii) construction, (iii) implementation, and (iv) evaluation. In more general terms, and adopting the framework of quality control, each of these stages corresponds to an action on the part of the syllabus designer, namely (a) Clarification of Tasks, also Needs Analysis, and Quality Functions Deployment (QFD) during the preparation and construction phases and (b) Failure Mode and Effects Analysis (FMEA) as part of the evaluative stage. A further step may follow evaluation and that is the Design of Experiment (DoE), intended as a way to control for errors and requiring an experimental setting. More in detail, QFD is a

> method to transform user demands into design quality, to deploy the functions forming quality, and to deploy methods for achieving the design quality into subsystems and component parts, and ultimately to specific elements of the manufacturing process. (Akao 1994, chap. 12)

On the other hand, FMEA and DoE are applied to eliminate errors through a statistical and probabilistic approach, i.e. with control factors and a quality index, to make predictions about future behaviour on the basis of how things have gone in the past. Let us consider each stage at a time.

## 3. Preparation and construction

The preparation stage usually includes needs analysis, with the purpose of clarifying tasks, and means analysis, i.e. the analysis of the setting and its constraints (cf. Alexander 2008, 82). The construction stage highlights the principles informing the EAP pedagogy and orient the instructor's choices towards one model or another, or a mixture of more.

Many authors, especially in the field of EAP, have stressed the importance of needs analysis because one is often dealing with very specific needs and therefore they must be identified as clearly as possible (see, for example, Spector-Cohen et al. 2001, 380; Alexander et al. 2008,

134; Liu et al. 2011), including the specific contexts and academic cultures of different subject areas. Needs analysis is aimed at narrowing down the language and skills that need to be taught. It tries to capture the target needs, what students need to be able to do as a result of the course (Target Situation Analysis) and the lacks, their capabilities now (Present Situation Analysis) and wishes, what learner wants to learn (cf. Flowerdew and Peacock 2001, 178; Hyland 2006, 196; Flowerdew 2013, 1906). This was elicited through a questionnaire (see Appendix) which showed that learners needed to be able to read academic texts, but wished to improve their conversation skills and were lacking grammatical accuracy. The questionnaire was drawn up by the course lecturer and administered following a language proficiency test assessing entrance level (expected B1), which was not done by the course lecturer, but rather developed and administered by another member of the teaching staff, the "Collaboratore and Esperto Linguistico", a NS teacher (what used to be called Lector), who offers general English classes on remedial grammar and vocabulary throughout the duration of the EAP course. Results of the entrance tests are usually quite disappointing, as the expected level is not met and the attested level is too low for the course. This means that the low scorers who attend the EAP course have to do the remedial grammar and vocabulary.

Seventy students responded to the questionnaire. The information collected via the students' questionnaires was complemented by informal interviews of colleagues in the target communities of practice and disciplines. The situation or domain which was identified is the 'academic' one. Students taking this English language course are third year undergraduates in their first semester studying a large variety of disciplines falling under the wider rubric of the Humanities. This implies that some, though not all, will have already chosen a topic for their final BA thesis and therefore need to read highly specialised texts in English on topics that are of direct interest to them. The bibliographies they deal with might contain references in English. Some other students, although not the majority of them, will have been exposed to texts in English as part of the curricular courses' reading lists. This is truer of certain subject matters, less of others. For example, Film Studies regularly includes Anglo-American set texts that are not available in Italian (as evident from the students' responses to questions no. 1 and 2 in the questionnaire), and so does Classical Studies, while Modern Philology, focusing on Italian language and literature, does not. These differences tend to affect students' motivation. Those who are required to read specialised literature in English as part of their research towards thesis writing or as preparation

for content-specific exam tend to see an EAP course as an opportunity to satisfy a tangible need. Utilitarian motives like this usually create high expectations of the course (as shown by the responses to question no. 2). If the course does not fail to meet such expectations, the outcome of the student's learning process is usually quite positive.

In terms of topics, or fields, the broad area that needs to be covered is the Humanities, which, however, encompasses a wide range of disciplinary fields, including Classics, Modern Philology, Film and Theatre Studies, Art History, Oriental Studies, Archaeology, Linguistics, Librarianship, etc. Despite being an opportunity for individualisation of academic subject, this clearly poses a problem of area-internal variation as it were (see also Hyland and Bondi 2006, 8 for a discussion of the term "discipline"; Hyland 2012a, 25 on disciplinary identities), which might be obfuscated by the little time available for enculturation into the literacy cultures of the various disciplines (irrespective of genre variation, a scholarly essay on contemporary American cinema will differ greatly in its rhetorical structure and linguistic choices from an essay on Cicero as a result of the different ways in which each academic and disciplinary community engages with the process of knowledge construction and negotiation). Although the Humanities share a common interest in

> the human condition, using methods that are primarily analytical, critical, or speculative, as distinguished from the mainly empirical approaches of the natural sciences (http://www.academicroom.com/topics/humanities-definition)

in actual fact they are found to differ greatly in their methods and how they organise knowledge. Following the OED definition of the Humanities,

> b. In pl. (usu. with *the*). The branch of learning concerned with human culture; the academic subjects collectively comprising this branch of learning, as history, literature, ancient and modern languages, law, philosophy, art, and music. Hence also in sing.: any one of these subjects.
> The humanities are typically distinguished from the social sciences in having a significant historical element, in the use of interpretation of texts and artefacts rather than experimental and quantitative methods, and in having an idiographic rather than nomothetic character.

Because of this and the specific national context, a non-English speaking country with English as a foreign language, thus a mostly monolingual class, reading was chosen as the core component of the syllabus, thought to be the skill more immediately of use to the course

participants (for whom the course is mandatory). Given the EFL context, it seemed necessary to place the same emphasis on both language and study skills, reading in particular (a similar scenario is described by Spector-Cohen et al. 2001).

For all the current effort by Italian universities to promote internationalization, classes remain mostly monolingual. Besides, still very few students from these cohorts are aiming to continue their studies abroad and therefore they do not need to improve their speaking, listening and writing skills for their academic studies. Also, Humanities students are 'good' readers in principle, or at least inclined toward reading as a result of their studies. However, although the main objective was to develop reading skills, listening and speaking/discussion activities were also planned as a way to alleviate the learning load, which otherwise would have been too much geared towards the written culture. A second reason was to take into account students' declared needs, who quite surprisingly, picked speaking as their priority (question no. 1), contradicting question no. 5 where reading was chosen by 67% of the group as the most important skill for their academic studies. However, Liu et al. (2011), following an experimental study of the items from almost a thousand questionnaires administered to six universities in Taiwan, conclude that students' needs are a "multiple and sometimes conflicting construct" (ibid., 276), as there are significant discrepancies especially in ESP/EAP contexts between what students want, i.e. their perceived needs, what they think is lacking vs. most necessary. In our study, while question no. 5 was meant to elicit the perceived necessities, no. 6 tried to get the skill they thought they were lacking and no. 1 the most desirable skills, i.e. what students want.

Based on the need to improve in their listening comprehension and note taking, English was chosen as the language of instruction. This content-based objective was parallel to the expected attainment level corresponding to B1+ reading (following the Common European Framework of Reference). One might question, however, that there is no real B1+ academic English for the Humanities when making use of authentic texts without any filtering for vocabulary (on vocabulary size and text comprehension see Nation and Waring 1997 and the more recent discussions in Cobb and Horst 2001 and Bruce 2011 both addressing the issue of the lexical threshold; cf. also Woodward-Kron 2008).

The preparation and construction phases also took into account responses related to the area of English most needed (question 4), vocabulary (rather than grammar), and to the text-types they would like to focus on during the course (question 13), both pointing to specialists texts

and to discipline-specific jargon, followed by newspaper articles and popularising prose. A few students also asked for fictional prose. The preference for newspaper articles seemed to contradict the perceived difficulties and rating of one's language ability (question 6), given that newspaper articles can contain a general language that requires advanced readers to comprehend (but see again Liu et al. 2011 on the discrepancies between wants and needs in EAP, and Miller 2011 questioning the efficacy of using non-academic reading texts in university-based intensive English programmes in the US).

Questions 8, 9, 10, and 11 were aimed at ascertaining the perceived attitude towards English also as a result of previous learning experiences. Out of the seventy answers, only 24% of the students declared to have had negative learning experiences mostly due to the poor quality of their teachers and lack of investment in language education on the part of the government (for example, by changing teacher every year). Learning styles were only marginally elicited through the questionnaire (see question 3).

The ranking of topic preferences (question 14) was extremely varied, reflecting the range of subjects in which these students can major. However, answers pointed to art history, theatre and cinema and literary criticism as the top choices, followed by classical studies, history and linguistics. Most respondents ranked philosophy and rhetoric last.

This analysis prompted a number of questions concerning genres and their conventions, rhetorical functions and their linguistic realization with a view to pinning down what is special to the Humanities in terms of phraseology, discipline-specific vocabulary and rhetorical organisation. To address questions such as "What are the language features special to the Humanities?", "Which texts and text-types best reflect such specificities?", "How can rhetorical functions and generic conventions be identified?" and "How can study skills be improved accordingly?", a mixed design method integrating corpus and genre research appeared as a viable resource.

In consideration of this scenario, neither a wide-angled (i.e. targeting a broad field) nor a narrow-angled (i.e. targeting one particular discipline) syllabus seemed to offer ideal feasibility, but rather what Basturkmen (2003, 50 ff.) calls a "type 3 syllabus", i.e. a syllabus that "present[s] an array of texts from the various sub-fields…a conglomeration of varieties". Following the distinctions reported in the literature (see Basturkmen 2010, 59; Jordan [1997] 2012, 61-63), a content-based syllabus was developed whereby the language, skills and academic conventions are associated with the students' subjects and the lexico-grammatical contents are taught as a part of whole texts, as encountered in the readings selected. In this sense,

the syllabus was not strictly a grammatical syllabus that moves from simple to complex. Rather, since the texts were the carriers of the features of language, the students were expected to learn, the grammatical structures were selected according to their frequency in the genres and text-types typical of a given discourse and estimated usefulness in relation to the study skill. Despite not being strictly communicative in approach either, the lexico-grammar was related to the purposes of communication or the functions of the various genres considered. As said, the syllabus was more topic- and content-based in angle in that the topics were selected among those suggested by students' responses and from their specialist studies and the vocabulary and grammar practised accordingly (on the benefit of subject-specific input into EAP courses, see Basturkmen 2003; Hyland 2006).

## 4. Implementation: *Reading Skills in English for the Humanities*

As a result of the preceding steps, a 36-hour course entitled *Reading Skills in English for the Humanities* was implemented aimed at improving reading and comprehension skills of humanities texts related in content to particular disciplines within the humanities. The syllabus is shown in detail in the Appendix. Each row contains the rhetorical and communicative functions of the text-types, the corresponding reading skills and lexico-grammar that need be developed in order to understand the texts and some examples from the language of the texts read in class. It can be read either from left to right or right to left without any difference, reflecting the design process described below switching between genreand corpus back and forth.

The following set of texts and genres was chosen and presented to the students in the following order: the Introduction from *Humanism* by Tony Davies; the Introduction from *Orientalism* by Edward Said; select essays from *The Artist's Reality: Philosophies of Art* by Mark Rothko; and the various texts from the *Gauguin Exhibition* booklet at Tate Modern, London. These were alternated with select readings and discussion activities from *English for the Humanities*, a textbook by Johannsen (2006), whose reading passages, while displaying a level of English lower than that of the authentic texts, did however offer a thematic linkage (cf. Spector-Cohen et al. 2001). The relevant syntactic and lexical features were then practised with Murphy's *English Grammar in Use*: *Intermediate*. (Reference is made in the syllabus to the various Units in Murphy's book to help students with remedial grammar).

From the point of view of authenticity, apart from the texts in the *English for the Humanities* textbook, the texts chosen were all authentic and likely to contain the language and structure of some of the texts that the students are asked to read in English for their final thesis (the exhibit material was meant more for the art history students, but also on the assumption that humanists in general enjoy going to exhibitions). From the perspective of genre theory (see Feak and Swales 2011), two instances of the same part-genre were chosen, namely the introductions to two important books such as *Humanism* and *Orientalism*. These were thought to provide students with insights into common generic features, while also lending themselves to an analysis of author-specific, stylistic differences within the discipline that can be broadly identified as intellectual history. Davies and Said have both been influential academic writers and in these books they deal with topics that are general enough to be of interest for all humanists without displaying vocabulary too specific to one discipline and thus constituting a deterring element for those learners who do not specialise in that one discipline. Rothko's essays were chosen for a number of reasons. First, the course instructor's own taste. Rarely is a painter also a remarkable writer capable of such appealing style. Second, they are representative of argumentative (and polemical) writing, which is one of the functions of academic discourse (see Nelson et al. 1987). Third, the topic, a philosophical reflection on art by the artist himself, which was thought to be interesting for everybody, not just the many art history students taking the course, and therefore facilitating the acquisition of vocabulary and syntax as well as the reading skill. Finally, compared to the other much more abstract readings, the Gauguin booklet was thought to offer the more concrete artistic terms, while also containing more general academic verbs and nouns (not just those linked to the exhibition experience).

The research article was left out on purpose to focus on genres that are not sofixed, but are loose enough to encompass the variety of language these learners need, in the belief that variety is what characterizes the Humanities. Genre theory thus guided the first choice of the texts and helped determine which rhetorical conventions are typical of a given academic genre across disciplines, giving precedence to whole texts over fragmented world of the concordance lines (cf. Swales 2002), and so that the EAP teacher can train students to do the same (cf. Swales and Feak [1994] 2012). However, because of the very nature of the Humanities, which defy any strict genre categorization, corpus linguistics can help pinpoint linguistic regularities and variation as attested in corpora, leading to a more accurate syllabus (see Flowerdew 1993, 2009). Together they

assist the course developer in delineating the "disciplinary textuality" (Fuller 1998, 47), i.e. the content and the language in which it is expressed.

## 5. Identifying the disciplinary textuality

This notion is variously detailed in the literature and its descriptive treatment oscillates between the identification of functions (whether rhetorical, communicative or linguistic, e.g. defending one's claims against criticism or abstraction) and the stress on language forms (e.g. linking words and connectivity) and it is not always easy to tell apart which is the outcome of the corpus and which of the genre approach. And yet, in line with Flowerdew and Peacock's (2001, 14) claim that "pride of place in EAP research must go to descriptions of target language that is appropriate to the target disciplines, occupations and activities", part of the implementation phase was aimed at describing the disciplinary textuality. The mixed-method approach to syllabus development was preceded by an overview of some available literature.

Fuller (1998) explores what is typical discourse in the sciences and the humanities and stresses the overlap between "the characteristically forbidding forms of academic science, such as high degrees of nominalisation, embedded causality, technical lexis and mathematical equations" (Fuller 1998, 35) and reasoning by abstraction, "moving from an instance or collection of instances, through generalisation to abstract interpretation" (ibid., 47). Similarly, Woodward-Kron (2008) mentions cause and effect logical relations, abstract terms such as *problem* and *argument* ("summary nouns", following Swales and Feak's [1994] 2012 nomenclature) among the typical features of EAP vocabulary in the humanities in place of the more technical jargon of the hard sciences. Bruce (2011) focuses on intra- and intersentential connectivity as a general feature of academic discourse. Bennett (2009) usefully compares style manuals for academic writing to conclude that there is continuity between the soft and hard sciences and a remarkable degree of consistency exists as regards the general principles and main features of academic discourse in English such as textual organisation, cohesion and coherence and the lexical and grammatical features to be used. Charles (2007, 2011) has added the corpus-informed observation that the learning input should encourage learners to see "the connection between general rhetorical purposes and specific lexico-grammatical choices" (Charles 2007, 289). Adopting a corpus methodology, Oakey (2002a, 2002b) has highlighted some frequent phraseology in different academic disciplines ranging from

social science to engineering, but does not describe disciplines within the humanities.

Swales and Feak's ([1994] 2012) contribute with "language foci" sections, providing a valid indication of the linguistic areas worth focusing upon. These include a mixture of detailed features such as linking words and phrases, *This* + summary word, and more general areas such as the grammar of definitions or reporting and attribution (shown to be constitutive of academic discourse in general). Among the more detailed features are: mid-position adverbs (e.g. *have recently been produced*), verbs and agents and the passive voice, clauses with *–ing* of result (e.g. *thereby causing*), indirect questions (*whether*…), linking *as*-clauses, qualifications (modal expressions like *it is likely*…), prepositions of time, *that*-clauses as direct object, evaluative adjectives across disciplines (*scholarly*, *sound*, *thin* vs. *rigorous*, *scholarly*, *anecdotal*), theme fronting (inversion). In the latest 2012 edition of their book, they have broadened the list of language foci to include nominalisation and complex noun phrases, known to be a feature of academic and specialised language, as observed by studies on readability and written academic discourse (cf. Miller 2011, 36 who examines the features employed by academic writers "to achieve information packaging efficiency, typically around the noun phrase") and by research into the role of the noun phrase and abstraction as a daunting feature of academic language (Woodward-Kron 2008; Parkinson and Musgrave 2014).

Nominalisation is also present in some pedagogical books on EAP, for example, Thaine (2012, Unit 4) and Cox and Hill (2011), where it is illustrated as a process where the final form is the nominalised one (see Cox and Hill 2011, 206-207):

> How people create a dictionary→ How a dictionary is created→ The process of dictionary creation

but not in the coursebook by Glendinning and Holmström (2004), for example. For this reason, the texts chosen for reading in class were analysed with special attention to the constitutive role of nominal sequences in the discourse of the Humanities. Starting from the Introduction to *Humanism*, various sequences with complex head nouns and prepositional phrases or relative clauses as post-modifiers were identified manually through linear reading of the text, e.g. *the secular and rational decencies of contemporary civilisation*; *the resulting accumulation and contest of meanings*; *the vertiginously proliferating and often contradictory senses assigned to the word*; *the uses to which the concept has been put in different times and situations*; *the often bitter*

*contentions in which it has taken on such an array of competing significations and values*.

The text is indeed interspersed with these kinds of sequences. However, what strikes the analyst about them is the lexical variety that tends to obfuscate any structural regularity, which instead should be made more visible to the researcher/course instructor first and foremost, for learners to be able to make sense of the syntactic complexity of nominalisation and ultimately understand the meanings construed in the text. This can best be evidenced by using corpus observation tools, that is by concordancing the text under consideration, here *Humanism*, and by contrasting it with another exemplar of the same part-genre, such as Said's Introduction to *Orientalism*, to see if any regularity at all can be identified as to the lexical choices contained in the nominal sequences *the…of*, *the…which*, *the … -ed*, and the way noun phrases are modified. Indeed, corpora can be used to capture variation within fixedness across styles, genres, domains, etc.

Therefore, a number of searches were carried out with a view to highlighting any relevant patterning. In the following figures (Figs. 1 and 2) the concordance lines for some of these searches are displayed for the purpose of illustrating the methodology followed. Figure 1 shows the concordances of the search for all words ending in **ion/*ions*, while figure 2 shows results from a search for the frame *the...of* with two intervening words (retrieved through the use of the wildcards).

1 e to identify it as the birthplace of the modern world.4 The resulting **accumulation** and contest of meanings
2 ts; and for anyone attempting to offer an account of those meanings, the **attraction** of Humpty Dumpty's
3 man freedom and dignity, standing alone and often outnumbered against the **battalions** of ignorance,
4 vely controversial. On one side, humanism is saluted as the philosophical **champion** of human freedom and
5 self-evidently for the secular and rational decencies of contemporary **civilisation** (i.e. of people like
6 the period from the fifteenth to the later eighteenth century, while the **Conclusion** goes back still further, to
7 , it carries, even in the most neutrally descriptive contexts, powerful **connotations**, positive or negative, of
8 nineteenth-century word) the 'Renaissance', a dauntingly complicated **constellation** of political, cultural and
9 what is 'at stake', historically and ideologically, in the often bitter **contentions** in which it has taken on such
10 meanings, the attraction of Humpty Dumpty's approach to the problems of **definition** is obvious. Life
11 Trust, on what you mean. The first problem, as always, is the problem of **definition**. So let's start with the
12 INTRODUCTION: TOWARDS A **DEFINITION** OF HUMANISM What is humanism? Well, that all
13 all see, is inseparable from the question of language. 'Man', in the old **definition**, is the 'talking animal'. The
14 g to Johnson's Dictionary, a humanist is 'a grammarian; a philologer', a **definition** that suggests how low

15 this book a good deal more straightforward, if I could simply set out my **definition** of humanism on the first
16 rescuing a single stable 'meaning', or even a range of sharply-focussed **definitions**; still less of suggesting,
17 ision making it all the more serviceable as a shibboleth of approval or **deprecation**. To some modern
18 ows after one accused the other's latest book of 'residual humanism', a **description** which was taken, rightly,
19 ts and debates in politics, science, aesthetics, philosophy, religion and **education**; and in spite of the
20 whose very existence is dismissed by some twentieth-century historians as a **fiction**, even while others
21 lism and scientific positivism that seemed to many to be undermining the **foundations** of Anglican belief.
22 atively offered by the Oxford English Dictionary in truth represent only a **fraction** of the senses and
23 from the tragedian Achaeus Eritrieus to the historian Zosimus. Known to **generations** of scholars and
24 connotations, positive or negative, of ideological allegiance, its very **imprecision** making it all the more
25 he foundations of Anglican belief. Humpty has been described as 'Verbal **Inspiration** sitting on a wall of
26 or the oppressive mystifications of modern society and culture, the **marginalisation** and oppression of the
27 ean languages was being done, in the nineteenth century, in Germany. The **motivation** behind the great
28 has been denounced as an ideological smokescreen for the oppressive **mystifications** of modern society
29 st what I choose it to mean. No such luck, alas. The seven distinct **sub-definitions** of humanism rather
30 **INTRODUCTION**: TOWARDS A DEFINITION OF HUMANISM What is humanism? Well, that all
31 carefully selected examples and a contemptuous disregard for any prosaic **objections**, that it means just
32 grammarian; a philologer', a definition that suggests how low that noble **occupation** had fallen by the later
33 ve mystifications of modern society and culture, the marginalisation and **oppression** of the multitudes of
34 y the Catholic-inspired Oxford Movement and the row precipitated by the **publication** of the doctrinally
35 ral authority, real, absent or desired, of those who use it. The important **question**, over and above what the
36 ngs', the philosophical egg goes straight to the heart of the matter: "The **question** is" said Humpty Dumpty,
37 ical antihumanisms (some of which will be explored in later chapters), the **question** of humanism remains
38 ological authority; and humanism, as we shall see, is inseparable from the **question** of language. 'Man', in e
39 s to which the concept has been put in different times and situations, the **questions** it has tried to answer,
40 Essays and Reviews (1860). Both of these, in their contrasting ways, were **reactions** to another kind of
41 ing key concepts and debates in politics, science, aesthetics, philosophy, **religion** and education; and in
42 helped to articulate all the major themes of the continuously unfolding **revolution** of modernity, structuring
43 Christ Church in the 1850s and 60s, was at the epicentre of the theological **ructions** caused by the Catholic-
44 htenment' essay from the educational reforms of Wilhelm von Humboldt, the **secession** of the North
45 itter contentions in which it has taken on such an array of competing **significations** and values. For the
46 e senses and contexts in which the word has been used, and a drastic **simplification** of those. It is one of
47 explore the uses to which the concept has been put in different times and **situations**, the questions it has

**Fig. 1. Concordances of **ion*/**ions* from the Introduction in *Humanism.***

The data in Figure 1 prompt a number of considerations. First, of the 47 instances of singular and plural nouns ending in *-ion*, only a few are repeated and these are *definition* and *question*. This repetition points to some broader functions of the text under analysis, namely the need to introduce the reader to the topic of the book (what it is about, i.e. which questions are going to be answered) and to the key concepts which are the object of investigation of the various chapters (as announced in the title of the Introduction, *Towards a Definition of Humanism*, also appearing in the concordance lines, see nos. 12 and 30). Through further sorting of the data displayed in the figure, collocates come to the fore such as *the question of humanism*, *the problem(s) of definition* pointing to the difficulty of defining complexity as it is encompassed by terms such as "humanism" (see also *sub-definitions*, line 29). This observation triggers a number of further searches on the part of the researcher/course instructor including a search for *isms* (which yielded 12 instances of *humanism*, but also *socialism*, *rationalism*, *realism*, *positivism*, and interestingly, *antihumanisms*), a search for language used to define terms, e.g. verb forms such as *mean(s)* and *is* (yielding, for example, *Humanism is a word with a very complex history*, *it is synonymous with…*, *a humanist is…*), and a search for what seems a productive structure, *the…of* with a varying number of intervening words (see Fig. 2 which captures two intervening words and the lexical material inside the prepositional phrase).

| | | | |
|---|---|---|---|
| 1 | tures through the looking-glass was Alice Liddell, | **the seven-year-old daughter of Henry George Liddell**, | ex-Head Master of Westminst |
| 2 | methodology and much of its material from | **the philological researches of Franz Passow**, | Professor of Greek at the Universit |
| 3 | nth century, in Germany. The motivation behind | **the great resurgence of German philological and archaeological scholarship** | was |
| 4 | was a reformed educational system inspired by | **the romantic hellenism of Winckelmann and Goethe**;1 | and the word the reformers in |
| 5 | lesiastical court, were acquitted on appeal by | **the 'King's men' of the Privy Council.**2 | But Humpty is also a philological despot |
| 6 | can mean 'a nice knock-down argument'. One of | **the oldest meanings of the word** | is 'exulting over the defeat of an enemy', as th |
| 7 | s to Essays and Reviews exulted, or gloried, in | **the humiliating 'knock-down' of their clerical persecutors.** | In short, he is, lik |
| 8 | roversial. On one side, humanism is saluted as | **the philosophical champion of human freedom and dignity,** | standing alone and ofte |
| 9 | ich we must look as the only bulwark against | **the materialistic 'anarchy' of contemporary society**.5 | On the other, it has been |
| 10 | n denounced as an ideological smokescreen for | **the oppressive mystifications of modern society and culture**, | the marginalisation |
| 11 | n one sense or other it has helped to articulate all | **the major themes of the continuously unfolding revolution of modernity**, | structur |
| 12 | ilosophy, religion and education; and in spite of | **the anachronistic crankiness of some contemporary 'humanist' movements,** | and the |
| 13 | while the Conclusion goes back still further, to | **the early etymology of the word 'human' and its uses in antiquity**- | must look |
| 14 | tes Immanuel Kant's 'Enlightenment' essay from | **the educational reforms of Wilhelm von Humboldt,** | the secession of the North Amer |

**Fig. 2. Concordances of *the* *of* from the Introduction in *Humanism*.**

For want of space the same search done on Said's Introduction to *Orientalism* is not displayed. However, compared to Davies' Introduction, the kind of pre-modification of head nouns found is not as rich, which is a sign of a less lexically dense and syntactically simpler prose. But other features were retrieved through an initial analysis of the wordlist and of the corresponding concordance lines, which make Said's writing no less complicated. For example, the use of complex prepositional phrases with *as* introducing Role found in the company of verbs such as *consider*, *identify*, *regard*, *take* etc. (as in *I study Orientalism as a dynamic exchange between individual authors and the large political concerns shaped by the three great empires...*) emerged as a typical feature of Said's introductory style and a sign of what Nelson et al. (1987, 248) call the hermeneutical *as*-structure of interpretation reflecting the tendency of the human sciences to allegorise concepts (that is, by taking something in the terms of something else). This in turn can be compared to a similar use of *as* in Davies, (see *humanism is saluted as the philosophical champion of human freedom*; *it has been denounced as an ideological smokescreen for the oppressive mystifications of modern society and culture*), in which the *as*-prepositional phrase is part of the language of definitions. As stressed by Swales and Feak ([1994] 2012), definitional information in Introductions is used to clarify terms that may be unfamiliar to the reader or controversial (as indeed are both *Orientalism* and *Humanism*).

The analysis of the concordance lines allows us to conclude that, as far as nominalisation is concerned, a lot of internal variation of the lexical material comprising it is characteristic. Nominalised sequences, at least in the texts considered, are lexically complex in that the kind of vocabulary that is packed into a noun phrase tends to be abstract and therefore more difficult for the learner to understand (e.g. *decencies*, *contest*, *contentions*). Furthermore, they are structurally complex as the degree of embedding and modification can get quite complicated with relative clauses in post-modification. Thus, understanding of the structure requires knowledge of other related grammatical contents such as reduced relative clauses and simpler clause constituents. Also, as a consequence of the richness it shows in humanistic texts, vocabulary building should be devoted attention and practice, and reading skills should be focused on unpacking information and understanding key terms. On the other hand, some major communicative function has been identified in the two introductions considered, namely definition, whereby concepts and terms are defined either because they are new or because they are ambiguous and an object of debate. This has confirmed the role of genres and part-genres as

codified ways that practitioners from different discourse communities employ to communicate specialised contents within the academia.

In sum, this combination of searches offers the analytical basis for a more global view of the genre of introductions and their function of defining and clarifying new disciplinary concepts. As Bondi and Diani (2009, 255) have argued, the phraseological analysis offers a way towards seeing the systematic relationship between language use and disciplinary practices. This is necessary for the course instructor to apprentice the students to their respective academic discourse communities, if EAP teaching is about the "deconstruction and reconstruction of academic texts and the discourses in which they are embedded" (Bruce 2011, 49).

Despite not being exhaustive, the method so far exemplified is meant to provide an illustration of how the EAP course instructor can make use of corpus tools to single out the vocabulary and grammatical resources which should be taught to help the students fare through a given semantic region. Combined with a genre-based approach to academic discourse, the information gained from corpus observations can be related to the rhetorical and communicative functions of a given text, genre and ultimately discourse and be used to advance students' understanding of academic communication. The various patterns that emerge by processing a text (or set of texts) through the concordancer and by comparing frequency distributions indicate what is typical (e.g. labelling reality, the language of definitions) and thus direct towards other texts that might be chosen to reinforce reading skills. In sum, the mixed approach herein proposed can help the course designer to perfect the design of the syllabus by adjusting the selection, grading and sequencing of contents, and by assisting in the choice of the relevant texts and activities aimed at developing reading skills (cf. Spector-Cohen 2001, 382 on the need for ongoing reassessment regarding the choice of specific texts).

In the specific case under evaluation, the order of presentation of the texts was reversed the following year, on the grounds that a movement from concrete to abstract would facilitate the learning process: the discourse of art exhibitions was the first piece of reading to which the students were exposed before moving onto more abstract reasoning as represented by Tony Davies and Edward Said's Introduction, advancing through the wonderfully argumentative and polemical style of Rothko's essay. The book review was added as a genre covering a variety of topics and more explicitly deploying the language of evaluation and engagement. Also, the methodology illustrated above assisted in the choice of the texts for the tests used for the evaluation.

## 6. Evaluation

The evaluative stage in the process of syllabus design answers the question "Have the needs and goals been met?" and does so by a systematic collection of information. In this instance, evaluation of the syllabus was done by the course designer, through collection and analysis of direct feedback from students and exams results.

A reading comprehension test was constructed *ad hoc* based on a text of the kinds read and analysed in class. The assumptions informing this choice are dual. On one side, the corpus has repetition and recurrence as its heuristics, on the other side, analogy and family resemblance (exemplars) are at the core of the notion of genre (however, Swales 1990, 46 places "the primary determinant of genre-membership on shared purpose rather than on similarity of form"). Thus, assigning a text for the test that has the same features as those read during the course draws from both approaches with the purpose of strengthening the learning process. Moreover, issues of test validity were also taken into account in that a test to be valid should only measure what it says it measures. By choosing a text of the type analysed in class it was possible to test what was covered by the syllabus and because of the format (a reading comprehension) the test also represented a simulation of study skills in use, i.e. reading (cf. Jordan [1997] 2012, 87), in line with the performance-based test approach (ibid.). Related to the overall course purpose of developing EAP reading skills, the test criterion was comprehending real world academic texts (cf. Bruce 2011, 202). In some testing sessions new extracts from exactly the same authors and books read in class were chosen (e.g. some of the chapters from *Humanism* on humanist printing, early humanism, etc.).

The achievement test is constructed as a reading comprehension with 15 questions of which 12 are open-ended and the remaining three are multiple-choice. The number of objective multiple-choice questions was limited to just 20% of the test so as to facilitate students' performance (e.g. *In paragraph x, Y's argument is that… A. B. C. D.*), while at the same time giving greater freedom in most of the answers. However, even the open-ended questions were guided towards a short answer thought to reduce production to a minimum. More in detail, in order to test comprehension, some questions are more analytical (e.g. *Find a noun in the text that means 'companion*'; *Give some examples of Greek myths mentioned in the article*) and aimed at testing knowledge of general academic vocabulary (*Pick two adjectives in the text that mean 'argumentative'*), others aimed at assessing global comprehension (e.g. *Say whether X is in agreement or disagreement with Y and give the line references for your answer*).

Occasionally, one question could test knowledge of technical jargon (e.g. art terms). Also, understanding of reference and coherence was tested as well as comprehension achieved through the variety of reading modes practiced with during the course, i.e. scanning for specific detail, skimming for gist and summarising the main point. On the whole, results were quite satisfying with an average pass rate of 80% and, limited to the testing session when mostly attending students sat the test, a bell-shaped distribution with 7% top scores and 10% fails. Findings of the evaluation show that, concerning the items in the lexico-grammatical syllabus, an especially problematic area appears to be that of linking phrases and logical connectives (such as *thus*, *although*, *as*, etc.). Except for *however* whose meaning was usually identified correctly as a result of its being very frequent in the texts read and tasks assigned during the course, the semantics of logical and cohesive relations proved to be sometimes obscure. Connection across sentence boundaries as exemplified by the connective *For* in sentence initial position was problematic even for the better scoring students. This might be due to the fact that although *For* was present in more than one text and consequently included in the syllabus, its function is not immediately salient to the intermediate learner, who thus requires more intensive exposure to its uses in context. Moreover, the students who had the most troubles were those who could not infer meaning from context or extract the relevant information by discarding irrelevant detail and those who did not prove to have enough mastery of reading strategies in terms of skimming through a passage quickly for essential information, or scanning for relevant detail. However, on average common phrasal verbs (*bring out*, *focus on*, *draw from*, etc.), typical evaluative adjectives (e.g. *influential*, *controversial*, *contentious*, *masterful*), word families and general academic as well as specialised vocabulary (e.g. *argue-argument-argumentation*, *paint-painting*, *draw-drawing*, *enlighten-enlightenment*) were correctly identified and understood and so were the macro communicative functions, a sign of successful reading ability.

## 7. Concluding remarks

In conclusion, no ideal syllabus is advanced here, but a proposal for a method of improving the design of EAP syllabi by harnessing corpus tools, especially concordances and word frequency lists, to drive teachers' choices of texts and language contents to be included in the syllabus. The argument has been made for the usefulness of integrating a genre-based, rhetorical approach to text analysis with the type of corpus investigation

that emphasises lexico-grammatical patterning. This has been argued to have direct pedagogical relevance in the context of English for the Humanities for a mixed-level group of EFL university students. From the perspective of quality control, a probabilistic approach has been taken to try and redress some of the failures associated with the proposed syllabus. A further step would be needed in the future, viz. an experiment comparing the learning outcome of two differently constructed EAP courses.

## References

Akao, Yoji. 1994. Recent approach of quality function deployment. In *QFD: The customer-driven approach to quality planning and deployment*, ed. Shigeru Mizuno and Yoji Akao, chap. 12. Tokyo: Asian Productivity Organization.

Alexander, Olwyn, Sue Argent, and Jenifer Spencer. 2008. *EAP essentials. A teacher's guide to principles and practice*. Reading: Garnet Publishing.

Basturkmen, Helen. 2003. Specificity and ESP course design. *RELC Journal* 34(1): 48-63.

—. 2010. *Developing courses in English for specific purposes*. London: Palgrave Macmillan.

Bennett, Karen. 2009. English academic style manuals: A survey. *Journal of English for Academic Purposes* 8(1): 43-54.

Bernardini, Silvia. 2004. Corpora in the classroom: An overview and some reflections on future developments. In *How to use corpora in language teaching*, ed. John Sinclair, 15-36. Amsterdam/Philadelphia: John Benjamins.

Bondi, Marina. 2008. Emphatics in academic discourse: Integrating corpus and discourse tools in the study of cross-disciplinary variation. In *Corpora and discourse: The challenges of different settings*, ed. Annelie Ädel and Randi Reppen, 31-55. Amsterdam/Philadelphia: John Benjamins.

Bondi, Marina, and Giuliana Diani. 2009. Linguistica dei corpora e EAP: Lingua, pratiche comunicative e contesto d'uso. *RILA* (*Rassegna Italiana di Linguistica Teorica e Applicata*) 1-2: 251-269.

Bruce, Ian. 2011. *Theory and concepts of English for academic purposes*. London: Palgrave Macmillan.

Charles, Maggie. 2007. Reconciling top-down and bottom-up approaches to graduate writing: Using a corpus to teach rhetorical functions. *Journal of English for Academic Purposes* 6(4): 289-302.

—. 2011. Using hands-on concordancing to teach rhetorical functions: Evaluations and implications for EAP writing classes. In *New trends in corpora and language learning*, ed. Ana Frankenberg-Garcia, Lynne Flowerdew and Guy Aston, 26-43. London: Bloomsbury.

—. 2012. "Proper vocabulary and juicy collocations": EAP students evaluate do-it-yourself corpus-building. *English for Specific Purposes* 31(2), 93-102.

—. 2014. Getting the corpus habit: EAP students' long-term use of personal corpora. *English for Specific Purposes* 35(1), 30-40.

Cobb, Tom, and Marlise Horst. 2001. Reading academic English: Carrying learners across the lexical threshold. In *Research perspectives on English for academic purposes*, ed. John Flowerdew and Matthew Peacock, 315-329. Cambridge: Cambridge University Press.

Cox, Kathy, and David Hill. 2011. *EAP now! English for academic purposes. Students' book. 2nd edition.* London: Pearson Longman.

Coxhead, Averil. 2000. A new academic word list. *TESOL Quarterly* 34, 213-238.

—. 2010. What can corpora tell us about English for academic purposes? In *The Routledge handbook of corpus linguistics*, ed. Anne O'Keeffe and Michael McCarthy, 458-470. London: Routledge.

Diani, Giuliana. 2012. Text and corpus work, EAP writing and language learners. In *Academic writing in second or foreign language: Issues and challenges facing ESL/EFL academic writers in higher education contexts*, ed. Ramona Tang, 45-66. London: Continuum.

Feak, Christine B. and John Swales. 2011. *Creating contexts. Writing introductions across genres*. Ann Arbor: The University of Michigan Press.

Flowerdew, John. 1993. Concordancing as a tool in course design. *System* 21(2): 231-244.

—. 2009. Corpora in language teaching. In *The handbook of language teaching*, ed. Michael H. Long and Catherine J. Doughty, 327-350. London: Wiley-Blackwell.

Flowerdew, John, and Matthew Peacock. 2001. The EAP curriculum: Issues, methods and challenges. In *Research perspectives on English for academic purposes*, ed. John Flowerdew and Matthew Peacock, 177-194. Cambridge: Cambridge University Press.

Flowerdew, Lynne. 2002. Corpus-based analyses in EAP. In *Academic discourse*, ed. John Flowerdew, 95-114. London: Longman.

—. 2004. The argument for using English specialized corpora to understand academic and professional language. In *Discourse in the*

*professions: Perspectives from corpus linguistics*, ed. Ulla Connor and Thomas Upton, 11-33. Amsterdam/Philadelphia: John Benjamins.
—. 2009. Applying corpus linguistics to pedagogy: A critical evaluation. *International Journal of Corpus Linguistics* 14(3), 393-417.
—. 2012. *Corpora and language education*. London: Palgrave Macmillan.
—. 2013. English for academic purposes. In *The encyclopedia of applied linguistics*, ed. Carol A. Chapelle, 1906-1912. Oxford: Wiley-Blackwell.
Fuller, Gillian. 1998. Cultivating science. In *Reading science. Critical and functional perspectives on discourses of science*, ed. Jim R. Martin and Robert Veel, 35-62. London: Routledge.
Gatto, Maristella. 2014. *Web as corpus. Theory and practice.* London: Bloomsbury.
Glendinning, Eric H., and Beverly Holmström. 2004. *Study reading. A course in reading skills for academic purposes*. Cambridge: Cambridge University Press.
Hyland, Ken. 2000. *Disciplinary discourses: Social interactions in academic writing*. London: Longman.
—. 2006. *English for academic purposes*. *An advanced resource book*. London: Routledge.
—. 2012a. *Disciplinary identities. Individuality and community in academic discourse*. Cambridge: Cambridge University Press.
—. 2012b. Corpora and academic discourse. In *Corpus applications in applied linguistics*, ed. Ken Hyland, Meng Huat Chau and Michael Handford, 30-46. London: Continuum.
Hyland, Ken, and Marina Bondi. 2006. Introduction. In *Academic discourse across disciplines*, ed. Ken Hyland and Marina Bondi, 7-14. Bern: Peter Lang.
Johannsen, Kristin L. 2006. *English for the humanities*. Boston, MA: Thomson Heinle Cengage Learning.
Johns, Ann M. 1997. *Text, role, and context: Developing academic literacies*. Cambridge: Cambridge University Press.
Jordan, Robert R. [1997] 2012. *English for academic purposes: A guide and resource book for teachers*. Cambridge: Cambridge University Press.
Krishnamurthy, Ramesh, and Iztok Kosem. 2007. Issues in creating a corpus for EAP pedagogy and research. *Journal of English for Academic Purposes* 6(4): 356-373.
Liu, Jin-Yu, Yu-Jung Chang, Fang-Ying Yang, and Yu-Chih Sun. 2011. "Is what I need what I want?" Reconceptualising college students'

needs in English courses for general and specific/academic purposes. *Journal of English for Academic Purposes* 10(4): 271-280.

Miller, Don. 2011. ESL reading textbooks vs. university textbooks: Are we giving our students the input they may need? *Journal of English for Academic Purposes* 10(1): 32-46.

Nation, Paul, and Robert Waring. 1997. Vocabulary size, text coverage and word lists. In *Vocabulary: Description, acquisition and pedagogy*, ed. Norbert Schmitt and Michael McCarthy, 6-19. Cambridge: Cambridge University Press.

Nelson, John S., Allan Megill, and Donald McCloskey. 1987. *The rhetoric of the human sciences: Language and argument in scholarship and public affairs*. Madison: University of Wisconsin Press.

Nesi, Hilary, and Sheena Gardner. 2012. *Genres across the disciplines: Student writing in higher education*. Cambridge: Cambridge University Press.

Newby, David. 2004. Syllabus and curriculum design. In *Routledge encyclopedia of language teaching and learning*, ed. Michael Byram, 590-594. London: Routledge.

Oakey, David. 2002a. Lexical phrases for teaching academic writing in English: Corpus evidence. In *Phrases and phraseology. Data and description*, ed. Stefania Nuccorini, 85-105. Bern: Peter Lang.

—. 2002b. Formulaic language in English academic writing. A corpus-based study of the formal and functional variation of a lexical phrase in different academic disciplines. In *Using corpora to explore linguistic variation*, ed. Randi Reppen, Susan M. Fitzmaurice and Douglas Biber, 111-129. Amsterdam/Philadelphia: John Benjamins.

Parkinson, Jean, and Jill Musgrave. 2014. Development of noun phrase complexity in the writing of English for Academic Purposes students. *Journal of English for Academic Purposes* 14(1): 48-59.

Peck McDonald, Susan. 2004. *Professional academic writing in the humanities and social sciences*. Carbondale: Southern Illinois University Press.

ReVelle, Jack B., John W. Moran, and Charles A. Cox. ed. 1998. *The QFD handbook. With a foreword by Dr. Yoji Akao.* New York: John Wiley.

Spector-Cohen, Elana, Michael Kirschner, and Carol Wexler. 2001. Designing EAP reading courses at the university level. *English for Specific Purposes* 20(4): 367-386.

Swales, John 1990. *Genre analysis. English in academic and research settings*. Cambridge: Cambridge University Press.

—. 2002. Integrated and fragmented worlds: EAP materials and corpus linguistics. In *Academic discourse*, ed. John Flowerdew, 150-164. London: Longman.

Swales, John, and Christine B. Feak. [1994] 2012 (3rd edition). *Academic writing for graduate students: A course for non-native speakers of English*. Ann Arbour: University of Michigan Press.

Thaine, Craig. 2012. *Cambridge academic English. An integrated skills course for EAP. Intermediate.* Cambridge: Cambridge University Press.

Thompson, Paul. 2006. Assessing the contribution of corpora to EAP practice. In *Motivation in learning language for specific and academic purposes*, ed. Zoe Kantaridou, Iris Papadopoulou and Ifigenia Mahili. Macedonia: University of Macedonia [CDROM].

—. 2007. Editorial: Corpus-based EAP pedagogy. Special issue of *Journal of English for Academic Purposes* 6(4): 285-288.

Thurstun, Jennifer, and Christopher N. Candlin. 1998. Concordancing and the teaching of the vocabulary of academic English. *English for Specific Purposes* 17(3): 267-280.

Woodward-Kron, Robyn. 2008. More than just jargon: The nature and role of specialist language in learning disciplinary knowledge. *Journal of English for Academic Purposes* 7(4): 234-249.

# Appendix: Questionnaire

*Fill out this questionnaire as accurately as you can.*

Major:

_____

| | |
|---|---|
| 1. | What tasks and activities will you be using English for? Circle the letter/s that correspond/s to your choice:<br><br>A. academic reading<br>B. (thesis) writing<br>C. speaking<br>D. listening |
| 2. | What do you expect from this course? |
| 3. | What would you like to learn/revise? |
| 4. | Which area of English do you think you need the most? Circle the letter that corresponds to your choice:<br><br>A. grammar<br>B. vocabulary |
| 5. | Rank the following skills in terms of their importance for your academic studies:<br><br>A. Writing<br>B. Speaking<br>C. Listening<br>D. Reading |
| 6. | Circle the number that best indicates your perception of your language ability in the major according to the following scale with 4 being the most:<br><br>Writing 4 3 2 1<br>Speaking 4 3 2 1<br>Listening 4 3 2 1<br>Reading 4 3 2 1 |
| 7. | How long have you been studying English? |
| 8. | How would you consider your previous learning experiences? Circle the letter that corresponds to your choice:<br>A. positive<br>B. negative |
| 9. | If your answer to no. 8. was B., then try to explain why. |

| | |
|---|---|
| 10. | What is your attitude to English now?<br>A. positive<br>B. negative |
| 11. | If your answer to no. 10. was B., then try to explain why. |
| 12. | What other languages have you studied? |
| 13. | Which text types would you like to be using in class? (e.g. essays and research articles, newspaper articles, specialist books, popularising prose, etc.) Give some: |
| 14. | Which disciplinary topics would you like to focus on during the course? Rank them according to your preference:<br>a. art history<br>b. classical studies<br>c. literary criticism<br>d. theatre and cinema<br>e. history<br>f. philosophy and rhetoric<br>g. linguistics<br>h. other (*specify:*_____) |
| 15. | How important do you consider certificates of language proficiency to be for your future career? Circle the letter that corresponds to your choice:<br>A. very important<br>B. quite important<br>C. not very important<br>D. they don't count at all |

## Syllabus

| ***Rhetorical and communicative functions of text-type*** | ***Reading Skills*** | ***Lexico-grammatical syllabus in context*** | ***Examples from texts*** |
|---|---|---|---|
| Introducing the reader to the topic of the book: definitions<br><br>Labelling reality and scientific entities | Understanding key terms<br><br>Understanding research questions<br><br>Unpacking information<br><br>Inferring the meaning of words | Word formation and word families (roots and affixes)<br><br>Vocabulary building: specialised jargon<br><br>Language used to define terms: *means*, *is* the present simple in academic English (Unit 2)<br><br>Articles and nouns, complex noun phrases and nominalisation (Units 79-80) | Nouns: *isms*, *scholars, academics, scholarship, tragedian, historian, humanists, orientalist, novelist, philosopher, grammarian, rhetoricians, translators, lexicographer, fellow artists, political theorists, politics, science, aesthetics, philosophy, religion and education, oriental studies, orthodoxy, theodicy, authority, tyranny, dignity, modernity, antiquity, development, constellation, mystifications, oppression, contest, contestant, theories, novels, accounts, civilization, specialization, etc.*<br><br>Adjectives: *historical (vs. historic), etymological, geographical, grammatical, morphological, philological and archaeological, biblical, theological, linguistic, clerical, controversial, materialistic, educational, eccentric, chronologicalcultural, colonial, imaginative, academic, methodological, ontological, epistemological, poetic, etc.* |

| | | | |
|---|---|---|---|
| | | | Lexical chains: a *reformed* educational system - the *reformers* - the educational *reforms*<br><br>Collocates: *critical debate, fleeting glimpses, superfluous detail, etc.* |
| Qualifying reality<br><br>Comparing and contrasting<br><br>Paraphrasing by using synonyms | Understanding evaluations<br><br>Understanding comparative language | Adjectives and adverbs<br><br>(Unit 98 on *-ing/-ed* adjectives; Unit 100 on adverbs and adjectives ending in *-ly*; Unit 101 on *good* vs. *well*, *fast*, *late*, *hard*, *hardly*, *what like? how?* etc.)<br><br>Comparison<br><br>(Units 105-108, Unit 107 p. 214 *as…as*; *so…as*; *…than*; Unit 108 p. 216 superlatives) | *Hum.*<br>Humpty-Dumpty smiled *contemptuously*, a *contemptuous* disregard, *pioneering* work, *doctrinally* unorthodox<br><br>*Orient.*<br>A French journalist wrote *regretfully*<br>*Historically and materially* defined<br>the *gutted* downtown area<br>There is a *much larger* number<br><br>*Hum.*<br>He is, *like Alice's father*, a lexicographer<br>Life would certainly be much easier...<br><br>*Artist's*<br>*Childish, irresponsible, stupid in everyday affairs, absentminded, rascal, etc.* |
| Time referencing and narrating | Understanding aspect in academic English: perfect vs. perfective | Present perfect simple and continuous, *for*/*since* (Units 7-12): a result now, a period until now, actions repeated over a period of time, etc. | *Hum.*<br>It *has helped* to articulate all the major themes…<br>Humpty *has been described* as…, his fall *has been associated*...<br>The truly pioneering work…*was being done* in the nineteenth century in G. |

| | | | |
|---|---|---|---|
| Talking about the past, narrating past events | Understanding narratives | Past tense: past simple (Units 5-6, 13-14), past perfect (Units 15-16),<br><br>*Used to* (Unit 18)<br><br>Appendix 2 p. 294 on past tenses | *Orient.*<br>Americans *have had* a long tradition of…; the Orient *has helped* to define Europe… It *had happened,* its time *was* over; Orientals… *had lived* there<br><br>*Gauguin*<br>*By now he had returned* to the South Seas... |
| Putting forward your point of view<br><br>Interacting and expressing opinions<br><br>Agreeing and disagreeing | Understanding argument building and writer's opinion<br><br>Understanding hedging | Modal verbs: [can, could, be able to (Units 26-27)], *must* expressing different levels of certainty / epistemic modality (Unit 28A, 28B), *may* and *might* expressing possibility (Unit 29A, 30 excluding 30D), *have to* (= it's necessary to do it, I'm obliged to do it) and *must* (Unit 31), *don't have to*, *don't need to* (Units 31C, 32), *should* expressing different degrees of obligation / deontic modality (Unit 33A advice) and epistemic modality (Unit 33B probability), *ought to* (Unit 33D)<br><br>had better (Unit 35A, 35B) expressing advice<br><br>Modal adverbs (see also Appendix 4 p. 296 on modals) | *Orient.*<br>*My contention is that…orientalism can be discussed and analysed as… My point is that…, It should be said at once that…, My argument, however, depends neither upon an…nor upon a…*<br><br>*Artist's*<br>*We must marvel at his wisdom…, how the artist might actually cultivate this...* |

| | | | |
|---|---|---|---|
| | | Offers and requests *Shall I…?Will you…?* (Unit 21D), requests, offers, permission and the imperative mood (Unit 37) | |
| Reporting and focusing on issues, processes and events | Understanding text organisation | Passive voice (Units 42-43)<br><br>Appendix 1 p. 293 on irregular verbs | *Hum.*<br>The truly pioneering work…*was being done* in the nineteenth century in G. Humpty *has been described* as…, his fall *has been associated*...<br><br>*Artist's*<br>He*'s held* to be…, they *are viewed* as…<br>Art *has often been described* as…<br>it *has been pointed out* that… |
| Establishing relations between parts of reality | Understanding phraseology | Vocabulary building: single word verbs vs. multiword (phrasal) verbs<br><br>(Introduction Unit 137, *leave out* 138C, *fill in* 138B, *write down* 142C 145) | Among the phrasal verbs encountered in the texts read and listened to are: *borrow from*, *draw from* / *draw upon*, *draw together*, *set out*, *set out to do sth*, *be associated with*, *look to*, *stand for*, *go back*, *turn back*, *fill in /fill out* (a form or questionnaire), *leave out*, *write down*, *hold on*, *look out / watch out*, *cut through*, *be immersed in, etc.* |
| Organising information 1 | Understanding the significance of references | Deixis: personal pronouns, reflexive pronouns (Unit 82)<br><br>Co-referential chains | *Hum.*<br>Like *them*, too, *it* carries…<br><br>*Orient.*<br>Now *it* was disappearing…<br>*I* shall be calling…<br>*It* is also the place…<br>…as *its* contrasting image |

| | | | |
|---|---|---|---|
| | | Existential predicates: *There is/there are* (Unit 84) | *Artist's*<br>*There is no* invention in him…, *we* deal in fantasies *ourselves*… |
| Organising information 2 | Analysing information in complex texts | Word order and clause structure<br><br>(Units 109-110, exercise 109.2 on p. 219)<br><br>Relative clauses (Units 92-95) | *Hum.*<br>The Lexicon (SUBJ.) borrowed (VERB) its descriptive methodology (OBJ.) from the philological researches of Franz Passow (PREPOSITIONAL PHRASE)<br><br>*Artist's*<br>This myth (SUBJ.) has (VERB) many reasonable foundations (OBJ.)<br><br>*Hum.*<br>His fall has been associated with the dismay occasioned in the Anglican faithful, *among whom the R. D. would have numbered himself,...*<br>…the complex of ideas *to which it referred*… |
| Organising information 3 | Identifying connections in text<br><br>Identifying main ideas and supporting information | Conjunctions, subordination, linking words and phrases (Units 113-116, 119-120)<br><br>*Nevertheless, however, although, though, indeed, in fact, moreover, in addition, in contrast, thus, furthermore, in spite of, despite, etc.* | *Hum.*<br>*For while* the language and the literature of the ancient Greeks continued to be studied (…), the truly pioneering work…<br><br>*Artist's*<br>*For, while* the authority of the doctor or plumber…, everyone deems himself…<br><br>*Gauguin*<br>…*but while* his fellow artists used painting to…, he argued that… |

**Notes**
The Units referred to are from Raymond Murphy's *English Grammar in Use*, 3rd edition, Cambridge University Press.
*Study Guide* on p. 326 and the additional exercises on p. 302 are recommended for self-study. The examples are all taken from the collection of texts read and analysed in class, which you can find in the black folder with the librarian. For their own practice, both attending and non-attending students are advised to fill in this grid with more pertinent examples from the same texts.

# CHAPTER THIRTEEN

# CHANGING THE BASES FOR ACADEMIC WORD LISTS

PAUL THOMPSON
UNIVERSITY OF BIRMINGHAM, UK

## 1. Introduction

In his book-length treatment of university language, Biber (2006, 3) observes that 'the description of vocabulary use in university contexts is an essential prerequisite to the development of effective teaching materials and approaches'. An important component of a comprehensive description of vocabulary use in academic settings is the word list (Nation and Webb 2011). Word lists can be used to judge which words learners are likely to encounter most frequently in target situations and, on the basis of this evidence, to decide which words should be concentrated on in teaching materials. The premise for this is that learners should focus on highly frequent items first and that in this way the time invested in learning will be maximally rewarded (Nation and Waring 1997). Word lists can be used by learners to set themselves learning goals and to decide which words to focus attention on, or they can be used by teachers and language testers for judging the content of reading or listening passages (Nation and Webb 2011).

The primary aim of this paper is to develop a word list for use in English for Academic Purposes (EAP) teaching and testing that focuses on the vocabulary needed by students who are preparing for listening to lectures in UK university environments: this word list will be called the 'Vocabulary for Academic Lecture Listening' word list (VALL). A secondary aim, which has to be achieved before the primary aim can be attempted, is to establish a robust basis on which to build the word list, and for this purpose the procedures by which word lists are developed will be reviewed and a set of alternative approaches will be evaluated.

The most influential EAP word list in recent years has been the Academic Word List (hereafter 'AWL') (Coxhead 2000); it is widely cited in EAP and related literature (as of March 2014, Coxhead (2000) has received 260 citations internationally, according to the Scopus database) and several textbooks such as Schmitt and Schmitt (2005) and Wells (2007), Wells and Valcourt (2008, 2010), have been developed around it.

The AWL contains 570 word families, beyond the 2000 most frequent word families in general English. These families were identified by Coxhead as having high levels of both frequency and range across a corpus of 3.5 million words of academic articles and textbooks, from a broad sampling of disciplines. To have range, word families have to occur in a wide variety of texts on different topics (Nation 1990, 13); in the case of the AWL, this means that the word family had to occur in texts from the four broad disciplinary domains represented in the corpus, and in at least 15 out of the 28 subject areas represented (Coxhead 2000). A word family is a headword and all its inflected and closely related derived forms (e.g. *green*, *greener*, *greenest*, *greenish, greenness*), and there are 3107 different types (unique word forms) in the 570 families. The frequency of AWL items was assessed by counting how often the types in a word family occurred within the corpus, after excluding the two thousand word families judged to be 'general English' (as distinct from 'academic English').

Coxhead (2002) used the General Service List (hereafter 'GSL') (West 1953) as the index of the 2000 most frequent word families in general language use. As Coxhead (2011) relates, the choice of the GSL was a controversial one because it was dated, but there was no comparable list available at the time that the Academic Word List was developed. As others have observed, highly frequent words such as *airport*, *television*, *video* are (predictably) missing in the GSL while archaisms such as *shilling* are included (see, for example, Hancioğlu, Neufeld and Eldridge 2008). The development in the last two decades of new corpora and new word lists based on frequency information derived from those corpora, however, means that it is now possible to reassess the value of basing academic word lists on the GSL.

Furthermore, as Coxhead (2011) emphasises, the AWL is representative of the vocabulary needed for reading written academic discourse, and it was not designed to describe spoken academic discourse. For these two reasons, it is therefore important, if we are to develop a new word list for listening to lectures and for assessing the vocabulary load of EAP teaching/testing listening material, that we should:

1. test the GSL against alternative word lists, and
2. use a corpus of transcripts of spoken academic discourse, rather than written language data, to derive frequency statistics.

The most frequent 2000 word families provide a much greater coverage of spoken discourse data than of written language data. In a classic study, Schonell, Meddleton and Shaw (1956) reported that the first two thousand word families provided 99% coverage of spoken discourse, but Adolphs and Schmitt (2003) found, in a study that used the CANCODE corpus as data, that the figure is approximately 95%. Biber (2006) indicated that lectures lie between conversation and academic writing in register features, and therefore one would expect to find that the most frequent word families will form a larger proportion of the vocabulary of academic lectures than they do of academic written discourse.

In this chapter, I will briefly describe the background and constitution of the GSL and the AWL before explaining the data used and the methods. I then test the combination of the AWL and the GSL on the academic lecture transcripts in the British Academic Spoken English (BASE) corpus,[1] using the same vocabulary profiling programme that Coxhead used: the Range programme (Heatley, Nation and Coxhead 2002).[2] Secondly, I evaluate different approaches to building an academic word list, using alternative 'general' word lists to identify the most frequent 2000 word families. The three alternatives to the GSL tested here are:

(1) the first and second thousand word families in the frequency lists derived from the British National Corpus (BNC) (Nation 2004)
(2) the first and second thousand word families in a composite list based on data from the British National Corpus and the Corpus of Contemporary American English (BNC/COCA) (Nation and Webb 2011)
(3) the Billuroğlu–Neufeld List (BNL), also known as the Bare Naked Lexis list (Neufeld and Billuroğlu 2006).

Once the best choice of 'most frequent 2000 word families' list has been made, a new 'Vocabulary for Academic Lecture Listening' word list will be presented. This provisional list is created, using frequency and range information drawn from the BASE corpus lecture transcripts, and it is then tested on another dataset, a set of lecture transcripts taken from the MICASE corpus, to check the coverage of the list, and to make

adjustments before finalising the list. The chapter then concludes with a discussion of pedagogical applications that exploit the new list.

## 2. The General Service List

The use of extensively researched word lists in language teaching has its origins in the Vocabulary Control Movement of the first half of the twentieth century. Gilner (2011) provides an excellent introduction to the history of this movement, leading up to the publication of the General Service List (West 1953). From this we learn that the GSL is a reissue of the "Interim Report on Vocabulary Selection" (Faucett et al. 1936), an annotated vocabulary list of about 2,000 words which developed out of two conferences in the nineteen thirties, sponsored by the Carnegie Corporation, and aimed at achieving consensus between the major vocabulary researchers worldwide.

It is important to note that the 2000 word families in the General Service List were not selected on frequency grounds alone. The purpose was to identify the words "considered suitable as the basis for learning English as a foreign language" (West 1953, vii) and additional criteria used for selection of words were: ease of learning (cost); necessity; cover; stylistic level; intensive/emotional words. Of these, the examples given for stylistic level choices are worth commenting on. West observed that 'personage' as a high stylistic device, and 'chap' and 'fellow' as colloquial (low stylistic) items do not need to be included as the word 'person' is sufficient in the early stages of learning. This criterion, however, may not be applicable in an EAP context, where learners need to be aware of what language they will be exposed to, which must be determined on the basis of the evidence that is available.

The GSL is no longer in print, but versions of the list can be found on the Internet, and there are differences between the versions: the Gilner and Morales version (www.sequencepublishing.com/academic.html), for example, contains 2303 word families, while Bauman (jbauman.com/aboutgsl.html) gives 2284, and the version that accompanies Heatley, Nation and Coxhead's (2002) Range programme contains 1992. In addition, while Nation's lists contain numbers, and the names of the months and the days, the other three, following the practice of the West publication, do not.

The University Word List (UWL) (Xue and Nation 1984; Nation 1990) was created in order to identify the vocabulary that a university level learner would need beyond the GSL, and it was important to include all items considered to be general, even if they were not specified in the GSL

(that is, the numbers, days and months). The UWL contains 836 word families and was the precursor to the AWL. The assumption underlying this work was that learners should first learn the word families in the GSL and then progress to the UWL. Nation and Waring (1997) report on studies which suggested that the GSL would give approximately 80% coverage of academic texts and the UWL an extra 10%.

The 570 word families of the Academic Word List were found to provide 10% coverage of a 3.5 million token corpus of academic texts, mainly textbooks and journal articles (Coxhead 2000) in addition to the 75% coverage provided by the GSL. Recent research has suggested that readers need between 95% and 98% for minimal to optimal comprehension of written text (Schmitt, Jiang and Grabe 2011), and that a vocabulary size of 4000 to 5000 word families is needed for 95% lexical coverage (Laufer and Ravenhorst-Kalovski 2010) and 8000-9000 for 98% coverage (Nation 2006) [these figures include proper nouns]. Van Zeeland and Schmitt (2013) report that little research has been done on the required lexical coverage of spoken language for L2 listening comprehension, and they propose that a figure of 95% coverage is sufficient for L2 language users to achieve satisfactory comprehension of spoken language input. The 95% coverage requires knowledge of 2,000 to 3,000 word families.

Although the AWL has become established as the standard source of information about written academic vocabulary, a number of criticisms have been levelled against it. Hyland and Tse (2007) question the existence of a core 'academic vocabulary' that students need to learn. They argue that there is no clear division between general and academic, and they also demonstrate that some highly frequent lexical items differ in meaning from one discipline to another. On this basis, they propose that vocabulary teaching should focus more on discipline-based needs. However, Eldridge (2008) in response points out that this is not always possible in EAP courses; for heterogeneous pre-entry teaching contexts he proposes that it is useful to teach a general vocabulary.

Hancuroğlu, Eldridge and Neufeld (2008) contest the claim that the GSL provides good coverage of text, observing that the 2K (second thousand) list of the GSL accounts for less than the complete AWL does (4.7% compared to 10%). As stated above, the GSL was not created solely on the criterion of frequency, but the weak performance of the GSL 2K suggests that an alternative 2K list is required, an observation also made by Nation (2006). While the first thousand words of the GSL seem to provide wide coverage, it appears that the second thousand word list does not and the aim of the following sections is to test alternative high-frequency lists to determine whether there is a better option available.

## 3. Data and methods

The British Academic Spoken English (BASE) corpus consists of the transcripts of 160 lectures and 39 seminars recorded at the Universities of Warwick and Reading in the period 2000-2005. The recordings are distributed across four broad disciplinary groups, as stated above, each represented by 40 lectures and 10 seminars (see Table 1). Only one lecturer is represented twice in the corpus, within the Social Sciences division – in the remaining 158 lectures, the lecturer is recorded in that lecture only.

For this study, only the lecture transcripts are used, and the total number of tokens in the dataset used comes to just over 1.1 million. The size of the corpus is small in comparison to the 3.5 million tokens in Coxhead's Academic Corpus but the high cost of transcribing large quantities of recorded speech means that good quality spoken language corpora are few in number. The balanced sampling frame used in the BASE corpus makes it the best available choice of corpus for this study.

| Disciplinary domain | Number of lecture transcripts | Size (in number of tokens) |
|---|---|---|
| Arts & Humanities | 40 | 295251 |
| Life Sciences | 40 | 281765 |
| Physical Sciences | 40 | 248282 |
| Social Sciences | 40 | 350799 |
| Total | 160 | 1176097 |

**Table 1. Composition of the BASE corpus lecture component.**

The transcripts are available in text and XML formats; here, the .txt files were used and they were pre-processed, with a series of search and replace operations (using regular expressions) as follows, in order to remove a number of 'noise' features in the data:

1. Speaker identifiers (eg, *nm0001*)
2. All truncated words
3. Filled pauses (for example, *er*, *mm*, *ah*).
4. Descriptions of non-verbal features such as [laugh] or [cough]
5. Repeated words, such as *the the the the*, were replaced with the first instance of the word.

The programmes used for this investigation were:

- the Range vocabulary profiling program (Heatley, Nation and Coxhead 2002) used by Averil Coxhead, and by Paul Nation, among others;
- text editing software (NoteTab Pro, NotePad++) for preprocessing the corpus files;
- Microsoft Excel, for storing and manipulating Range output.

The Range programme calculates the frequency and range of types and tokens in a given set of text files and also presents frequency and range information for the word families. Range is determined by calculating frequency of occurrence for each file. For this analysis, the BASE lecture files were merged into four files, one for each of the four disciplinary domains, rather than one for each lecture, as the aim was to ensure that there was an adequate range of vocabulary items across the four domains. As each file contained 40 lectures, I changed the requirement that there should be at least one occurrence per file; instead, I decided that the minimum required frequency for a file was eight occurrences.[3]

## 4. First stage of evaluation: The GSL and AWL

To evaluate the coverage of the GSL + AWL combination, the Range programme was run on the BASE lecture corpus using the GSL 1K and 2K and the AWL baseword lists; the results are shown in Table 2, with the equivalent figures given for Coxhead's (2000) analyses of her written language corpus.

It is important to note that a very high proportion of tokens in both analyses are accounted for by the first thousand word families. As noted in section 1 above, Biber (2006) has shown that lectures fall somewhere between conversation and academic writing in register features, and one would expect to find features of both conversation and academic writing in the vocabulary of academic lectures, and therefore a broader coverage by the first thousand words of the GSL. However, while the GSL 1K is 11.97% higher for the BASE corpus (spoken language) data than for Coxhead's written language corpus, the figures for the GSL 2K show a slightly reduced proportion. In other words, the most highly frequent words in general English (as represented in the GSL) occur far more frequently in academic lectures but the word families in GSL 2K do not. In fact, the 570 word families in the AWL provide wider coverage (4.84%) than the GSL 2K (4.13%). As Neufeld and Billuroğlu (2006) and Nation (2006) have observed, this suggests that the GSL 2K list is not an accurate representation of word family frequencies in modern day English, as one

would expect the third band of word families to account for less than the second. A further possibility is that, if the GSL were replaced by a new word list based on calculations of word frequency in a modern general corpus, then many of the AWL word families that make up that 4.84% would be placed within the most frequent two thousand words, a point that we will return to later in the next section.

| | Coverage of tokens: BASE lectures | Coverage of tokens: Coxhead 2000 study | Families found in corpus: BASE lectures | Families found in corpus: Coxhead 2000 |
|---|---|---|---|---|
| GSL 1K | 83.37 % | 71.4 % | 998 | 1000 |
| GSL 2K | 4.13 % | 4.7 % | 953 | 968 |
| AWL | 4.84 % | 10.0 % | 569 | 570 |
| Not in the lists | 7.66 % | 13.9 % | n/a | n/a |

**Table 2. Coverage of the tokens in the BASE corpus lecture transcripts using the GSL + AWL baseword lists in Range, with comparison to the coverage for Coxhead's (2000) written language corpus. The final two columns indicate the number of word families from each baseword list found in the data.**

## 5. Second stage of evaluation: Alternative sources of information about the first 2K frequent word families

Having determined that the combination of the GSL + AWL provides 87.5% + 4.84% coverage, we now move on to evaluate alternatives. We are looking to see whether other published lists of highly frequent 'general' vocabulary can provide better coverage than the GSL. For the evaluation, three other lists are used: the BNC, the BNC/COCA and the BNL lists.

The BNC frequency lists were created by Paul Nation using information from the BNC word frequency lists (Leech, Rayson and Wilson 2001) derived from the 100 million word British National Corpus. Unlike the GSL, no words are excluded on the grounds of stylistic level: both *chap* and *bloke* are in the BNC 1K list, as well as *person*. As Nation has noted, these lists are biased towards written language because the BNC contains a much higher proportion of written language data (90%) than spoken language (10%).

The BNC/COCA lists combine frequency information derived from two corpora: the British National Corpus (BNC) and the Corpus of

Contemporary American (COCA) English. The lists were developed by Paul Nation for a variety of purposes, including providing a reference point for establishing what vocabulary to use in graded readers, and also as a reference on vocabulary for language testing (Nation and Webb 2011). However, it should be noted that the 1K and 2K lists were specially developed by Paul Nation to include better representation of spoken language and of child language in particular, including lexis such as *angel*, *castle*, *doll*, *gorgeous* and *mama*.

The BNL list was created by Neufeld and Billuroğlu at Eastern Mediterranean University in Cyprus to form an integrated alternative to the AWL and the GSL (Billuroğlu and Neufeld 2007). The complete list contains 2709 word families. These word families were identified by taking a number of commonly used word family frequency lists, including the AWL, the GSL, and the BNC frequency lists, and then building a set of candidate items for their new list out of the sharedness of items across the lists. They found that the first five thousand word families for the BNC covered most of the items in the lists that they used and that the word families from GSL and AWL were found in all five 'one thousand' word bands. They then test the word family lists on a corpus of written texts (assembled for pedagogical purposes) and create a list of 2709 families, divided into six (unequal) sublists; for this evaluation, I took the first four sublists, containing a total of 2138 word families to obtain a set of as close to 2000 word families as possible. The BNL absorbs many AWL word families and the combined first four sub-lists (investigated here) contain 312 word families from the AWL.

The BASE corpus files were analysed using the different base word lists in Range and the results are shown in Table 3. For brevity's sake, the results are shown for the token coverage only.

| | GSL | BNC | BNC/COCA | BNL |
|---|---|---|---|---|
| 1K | 83.37% | 85.68% | 82.04% | First 4 lists combined |
| 2K | 4.13% | 5.65% | 6.5% | |
| 1K + 2K | 87.50% | 91.33% | 88.54% | 89.91% |

**Table 3. Coverage of the tokens in the BASE corpus, using the GSL, the BNC, the BNC/COCA or the BNL as the first two thousand word family lists.**

At 91.33%, we can see that the BNC list provides the best coverage of the four alternatives, and 3.83% more coverage than the GSL. How can we account for the differences in coverage between BNC and the GSL? They derive primarily from the composition of the text corpora used for

compiling the lists. The BNC contains a large amount of newspaper text and therefore gives a broad representation of social institutions and roles. This is evident in words in BNC 2K that refer to political groupings (*labour*, *conservative*, *liberal*, *democrat*), countries (*Australia*, *Britain*, *France*, *Germany*), social institutions (*Parliament*), social roles (*landlord*, *deputy*, *mayor*) and new technologies (*television*).

The GSL by contrast contains a range of items related to categories such as:

- emotions: *anger*, *anxiety*, *sympathy*
- materials for clothing and furnishing: *cotton*, *elastic*, *straw*, *wool*
- 'small town' occupations: *barber*, *merchant*, *priest*, *soldier*
- creatures: *elephant*, *goat*, *monkey*, *rabbit*, *pet, beak*
- house and farm tools: *plough*, *scissors*, *spade*

On the evidence of these word families, one can deduce that the GSL is composed predominantly of words from the world of fictional literature, where descriptions abound of social interactions, of human emotions, of creatures, of the materials and objects that surround us in our daily lives. For educational contexts such as university lectures, the BNC lists provide a much better general coverage because the BNC contains more words dealing with a society of work and politics, and with social concepts (in addition to words from literary texts). In EAP higher education teaching contexts where students are young adults or older, the BNC lists appear to offer a better guide to the high frequency vocabulary that learners could be expected to have knowledge of, as a basis for functioning in such educational contexts.

Does this suggest that the GSL is redundant? Here we have only evaluated the GSL in relation to the corpus of academic lectures and so it is not possible to claim that the GSL is no longer relevant as an indication of general language. On the basis of my analyses, however, I would argue that the BNC frequency lists are much better than the GSL for the purposes of profiling vocabulary in spoken academic discourse.

## 6. Vocabulary for Academic Lecture Listening (VALL)

### 6.1. Developing a new spoken academic list

Having identified the BNC 1K and 2K lists as the best available basis for building a Vocabulary for Academic Lecture Listening (VALL) word list, the next step was to determine the word families outside the top 2000

that have reasonable levels of both frequency and range. As explained above, the range score for this study was based on whether the tally of word family frequencies for any of the four domains was greater than seven. For a word family to have a range score of 4 (maximum), the word family would have to appear at least eight times in each category. All word families that achieved a range score of 4 were immediately added to the list, on the grounds that they have full range across all four domains. Families having a range score of 3, and a token frequency score overall of 40 or above, and those having a range score of 2, and a token frequency score overall of 75 or above were also added. In total, 193 word families were identified as candidates for the list, the majority coming from the 3K and 4K bands of the BNC lists (that is, in the 2001-3000 and 3001-4000 frequency).

| WORD LIST | TOKENS as % | TYPES as % | FAMILIES |
|---|---|---|---|
| BNC 1K (first thousand) | 85.68 | 15.10 | 999 |
| 2 (second thousand) | 5.65 | 12.28 | 993 |
| BNC 3K | 0.64 | 1.06 | 78 |
| BNC 4K | 0.67 | 1.10 | 74 |
| BNC 5K | 0.18 | 0.31 | 26 |
| BNC 6K | 0.08 | 0.12 | 7 |
| BNC 7K | 0.03 | 0.03 | 3 |
| BNC 8K | 0.02 | 0.04 | 3 |
| BNC 9K | 0.01 | 0.01 | 1 |
| BNC 10K | 0.00 | 0.01 | 1 |
| not in the lists | 7.05 | 69.93 | N/A |
| Sum | 1205513 | 27821 | |

**Table 4. The Range programme output using the 193 high range word families on the BASE lecture corpus [word lists 3-10 contain only the VALL word family candidates that are in that range].**

One item that was added was *so-called* which is not in any of the BNC lists but which can be treated as a distinct lexical item (Coxhead also included it in the AWL). With a range score of 4 and a frequency of 81, *so-called* is worthy of inclusion as a separate item, and was added to the list of VALL word family candidates.

For each of the lists 3 to 10, all non-qualifying word families were removed (word families that did not fulfil the frequency and range criteria), which left only the contenders for the VALL list, and then the Range programme was run again on the BASE corpus data, with the results shown in Table 4. The 193 word families that have been selected

provide 1.52% coverage of the tokens in the BASE lecture corpus. However, before this list and the statistics around it can be finalised, it needs to be tested on another dataset, and this is the subject of the next section.

## 6.2. Trialling the new lists on MICASE lectures

The set of 193 word families that have both frequency and range in the BASE data was then tested for coverage (in association with the BNC 1K and 2K lists) on a comparable dataset. For this purpose, the MICASE corpus (Simpson et al. 2002) was used, because it contains a wide variety of speech events including 62 lectures. MICASE does not contain a similar proportion of lectures across its four disciplinary domains in the way that the BASE corpus does, and it also features American English rather than British. However, the MICASE files present the largest available alternative collection of English language lecture transcripts. To prepare the 62 files for testing, the lecture transcripts were processed in the same way as the data for the BASE corpus (as described above), with all word repetitions, truncations, pauses and mark-up removed. The cleaned up text files were then run through the Range programme with an amended set of baseword lists. The BNC 1K and 2K lists were used and the 3-10 lists were the set of 193 word families identified; no other word lists were used.

As can be seen in Table 5, the coverage of the word lists is slightly lower (1.20% as compared to 1.52%) than it was for the BASE corpus, but the coverage of the BNC 1K and 2K lists is higher (91.72% as compared to 91.33%).

While nearly all of the first thousand word families appear in both sets of data, the number of word families in the second thousand that appear in the MICASE data is lower than for the BASE corpus (957:993). The main reason for this is that the word lists used with the Range programme are derived from the British National Corpus and there are several lexical items that are typical of British English and not of American English. For example, amongst the words which appear in the BASE corpus and not in the MICASE corpus are *cheque* (*check* in American spelling), *bugger*, *fetch*, *fortnight*, *petrol*, *pub*, *rubbish*, *trousers*. These differences serve to remind us that there are regional as well as genre and register variations and that word lists based on corpus-derived frequency information will not be applicable to all contexts.

| WORD LIST | TOKENS as % | TYPES as % | FAMILIES |
|---|---|---|---|
| BNC 1K (first thousand) | 86.31 | 19.93 | 993 |
| BNC 2K (second thousand) | 5.41 | 15.56 | 957 |
| BNC 3K | 0.51 | 1.40 | 77 |
| BNC 4K | 0.44 | 1.29 | 72 |
| BNC 5K | 0.15 | 0.40 | 26 |
| BNC 6K | 0.07 | 0.13 | 7 |
| BNC 7K | 0.00 | 0.03 | 3 |
| BNC 8K | 0.01 | 0.04 | 3 |
| BNC 9K | 0.01 | 0.02 | 1 |
| BNC 10K | 0.00 | 0.01 | 1 |
| not in the lists | 7.08 | 61.20 | N/A |
| Sum | 591792 | 18220 | |

**Table 5. The Range programme output using the 193 high range word families on the MICASE lecture corpus [word lists 3-10 contain only the VALL word family candidates that are in that range].**

The words which were treated as 'outside the lists' in the MICASE data were then examined and items which were both frequent and had a range of 20% across the 62 files were identified. The criterion of range (as it had in the investigation of the BASE data) was used to identify 8 new word families to add to the list: *cancer*, *decrease*, *index*, *layer*, *multiple*, *organism*, *protein*, *web*. At the same time, one word family was removed from the original lists that hardly occurred in the MICASE data: *wheat*.

The modified word list, consisting of 200 word families (see Appendix), was then tested on the BASE corpus with its new additions, and a further test was run on the MICASE corpus. The results of this test are shown in Table 6.

| | BASE | MICASE |
|---|---|---|
| BNC 2K | 91.33 | 91.72 |
| VALL 200 | 1.66 | 1.38 |
| *BNC 2K + VALL 200* | *92.99* | *93.10* |
| Off lists | 7.01 | 6.90 |

**Table 6. The Range programme output using the final 200 high range word families on the BASE lecture corpus and on the MICASE lecture corpus.**

From this one can see that the BNC 2K and VALL list provide 93% coverage of the tokens in the BASE lectures and 93.1% for the MICASE lectures in the two corpora of lecture transcripts. This is an impressive performance. At the same time, one should observe that the coverage provided by the 200 word list is relatively small, ranging between 1.38% and 1.66%. The Academic Word List, by contrast, provides 10% coverage. Is one justified in claiming that the 200 word list is worth spending time on? The first point to be made is that the first 2000 word families are providing far greater coverage of the spoken data than of the written data in Coxhead (2000), and therefore there is much less work for other words to do. Furthermore, the BNC 1K and 2K lists actually contain many of the family lists in the Academic Word List – 82 of the word families from the AWL appear in the BNC 1K list, while the BNC 2K list contains 206, meaning that 288 of the AWL families are contained within the BNC 1K and 2K.

Secondly, if we look at the number of tokens not included in the first 2000 word families and judge what proportion of those tokens are covered by the 200 word families then we can see that the VALL word families are providing a relatively high coverage. The off list category includes a large number of proper nouns (e.g., *Aristotle*, *Hitler*) and these should not be regarded as vocabulary items for teaching purposes.

Table 7 adds to Table 6 a further level of detail in which the names have been calculated separately from the rest of the off list items, and in the final row of the table, the VALL 200 coverage is shown as a percentage of all of the coverage by the words beyond the first 2000 – with a range between 18.7% and 21.0%. In other words, the VALL list provides coverage of approximately one in five words, when items from the first 2000, plus all proper nouns, are excluded.

| | BASE | MICASE |
|---|---|---|
| BNC 2K | 91.33 | 91.72 |
| VALL 200 (2) | 1.66 | 1.38 |
| Names | 0.77 | 0.90 |
| Off lists (4) | 6.24 | 6.00 |
| VALL 200 as percentage of 2 + 4 | 21.0% | 18.7% |

**Table 7. The Range programme output using the final 200 high range word families on the BASE lecture corpus, with proper nouns separated from the other off list items.**

## 6.3. Final list

The final list, then, is composed of 200 word families, in addition to the first 2000 word families which are in the Nation BNC 1K and 2K lists. The complete dataset has been created as an Excel spreadsheet which is available at www.birmingham.ac.uk/staff/profiles/elal/thompson-paul.aspx. The lists are additionally provided with information about the frequency for each word type in each of the four broad disciplinary domains in the BASE corpus so that users can see whether a given word family or type occurs with frequency in all domains or tend to be particularly frequent in one domain.

Seven families in the 2K list do not actually occur in the lecture data at all, and some do not occur with range. It would be possible to go a step further and remove the 2K word families which have low frequency/range in the lecture data altogether. This would counteract any tendency to regard the first 2000 as "general" vocabulary and the subsequent 200 word families as "academic" and it would then be possible to create a single high frequency lecture word list. Rather than prejudge the issue, however, by removing the low frequency/range 2K list word families at this stage, the solution I have taken is to leave those families in the final list. A justification for this decision is that the BNC 2K lists need to be tested on other academic corpus data, including written academic data, such as the British Academic Written English corpus (Nesi and Gardner 2012).

What is clear from these investigations is that the first 2200 word families (as identified in this study) feature very prominently in the language used in academic lectures providing 93% coverage of the tokens in the BASE corpus data. There is a danger of assuming that language learners who are preparing for university lectures already know the first 2000 word families but the more frequent the lexical item the more polysemous it is likely to be (Zipf 1945), which can mean that there are more meanings to learn for those items (although they may relate to the same conceptual structure, cf. Murphy, 2004) and the item is also likely to feature in recurrent phraseologies.

The observation by COBUILD researchers that the most frequent 2500 words account for 80% of English text led Sinclair and Renouf (1988) and Willis (1990) to propose the lexical syllabus, in which learning focused on the most frequent items and what they do. While not suggesting that EAP teaching which focuses on listening to lectures should similarly be wholly structured around a lexical syllabus, I would argue, however, that if 93% of the tokens occurring in the BASE lecture corpus are accounted for by 2200 word families then learners are likely to derive high rewards for the time and effort invested in learning this vocabulary.

## 7. Corpus investigation activities to use with VALL

In the process of building the VALL word list, we saw that the BNC 1K (in particular) and the BNC 2K lists provide an immense amount of coverage of the words used in academic lectures and we also noted the importance of seeing how words are used in context – words not inherently 'academic' *per se*, but they become academic through context as Hancuroğlu, Eldridge and Neufeld (2008, 463) have argued. This is a point also made by Hunston (2002):

> What makes text 'academic', then, is not the occurrence in isolation of certain specific items, but the ways in which certain items 'collocate' and 'colligate', in other words, the ways lexical items co-occur with other lexical and grammatical items (Hunston 2002, 12-13).

From these observations, we can propose that learners need, firstly, to give attention not only to the words in the VALL list but also look at the words in the BNC 1K and 2K lists (Paul Nation has repeatedly emphasised the importance of teaching these high frequency vocabulary items, cf, for example, Newton and Nation 1997) and, secondly, that these words should be studied in the contexts in which they typically occur in academic lectures. If this is the case, then, it makes sense to use either authentic recordings or transcripts of authentic lectures in teaching materials for academic lecture listening. The BASE corpus provides a wealth of evidence that learners can explore. The lecture corpus can be accessed online through Sketch Engine at the.sketchengine.co.uk/open/. This interface allows users to run concordance searches, sort data, obtain frequency information and read examples in context.

The BASE corpus home page allows the user to make a simple query rather like a Google search. The user can also choose to run the query on lectures from a particular discipline, or from one or more of the four disciplinary domains.

The focus of activities should be on the relationship of words to other words, the patterns they appear in, and the functions and contexts of these patterns. As an example, learners can carry out investigations (working in groups, for example) of how highly frequent lexical items are used. Table 8 shows the 30 most frequent verbs in the BASE corpus. The first three verbs are best ignored in this activity because they are so highly frequent, but for the verbs below them on the list one can sample the data so that the quantity is not overwhelming.

| Rank | Verb | Raw frequency | % | Rank | Verb | Raw Frequency | % |
|---|---|---|---|---|---|---|---|
| 1 | be | 69963 | 26.9 | 17 | use | 2018 | 0.8 |
| 2 | have | 18276 | 7.0 | 18 | give | 1914 | 0.7 |
| 3 | do | 11868 | 4.6 | 19 | mean | 1907 | 0.7 |
| 4 | go | 6711 | 2.6 | 20 | take | 1864 | 0.7 |
| 5 | get | 6503 | 2.5 | 21 | talk | 1622 | 0.6 |
| 6 | can | 6253 | 2.4 | 22 | could | 1573 | 0.6 |
| 7 | will | 5605 | 2.2 | 23 | need | 1385 | 0.5 |
| 8 | say | 4878 | 1.9 | 24 | might | 1358 | 0.5 |
| 9 | know | 4172 | 1.6 | 25 | call | 1294 | 0.5 |
| 10 | think | 3637 | 1.4 | 26 | work | 1218 | 0.5 |
| 11 | would | 3457 | 1.3 | 27 | put | 1215 | 0.5 |
| 12 | look | 2963 | 1.1 | 28 | try | 1194 | 0.5 |
| 13 | see | 2864 | 1.1 | 29 | find | 1167 | 0.5 |
| 14 | make | 2184 | 0.8 | 30 | should | 1140 | 0.4 |
| 15 | come | 2146 | 0.8 | 31 | happen | 1122 | 0.4 |
| 16 | want | 2119 | 0.8 | 32 | let | 1104 | 0.4 |

**Table 8. The 32 most frequent verbs in the BASE lecture corpus.**

Questions to investigate include:

- What nouns appear to the left of the verb?
- What nouns appear after the verb?
- Does the verb tend to be used in the active or in the passive voice?
- Which words occur most frequently immediately to the left or immediately to the right of the verb?

A search for *happen*, for example, shows that the verb occurs 1122 number of times in the corpus, and by choosing to view frequency information by 'node form' (see below), learners can see that 'happens' is the most common form of the lemma.

In the Sketch Engine frequency display, the letters 'p' and 'n' appear to the left of the word form, where 'p' stands for 'positive' and 'n' stands for 'negative'. Clicking on 'p' retrieves all concordance lines which include 'happens' (clicking on 'n' would retrieve all concordance lines without 'happens').

**Fig. 1. Screenshot of frequency information for forms of the lemma *happen*, taken from Sketch Engine Open.**

Sorting the data by first word to the left shows that the most frequent word directly to the left is the word 'what' and to the right we get 'to', 'when', 'if', 'is' and 'in'. 'Happens' is more frequent in the Life and Physical Sciences, while 'happened' is more common in Arts & Humanities and the Social Sciences. 'What happens' is used in scientific experiments to investigate cause and effect ('let's see what happens'; 'what happens is') and to formulate observations, while 'happened' is used of changes and events in the past, either past simple or present perfect. Interestingly, 'happen' does not have a negative semantic prosody in many of these instances, in contrast to what Sinclair (1991) and Adolphs (2006) both observe, which supports Tribble's (2000) notion of 'local' prosodies – semantic prosodies are related to genre, rather being universal. Sinclair made his observations on the evidence of a corpus predominantly composed of news texts and, as we know, newspapers focus on bad things that happen; our observations are based on the evidence of lectures, a different genre.

Investigations such as this can be set up as small group research projects where each group has a verb to investigate and the outcome of the activity is a group presentation to the class on the findings. Similar explorations can be done with adjectives, or nouns. For instance, Hyland and Tse (2007) comment on the typicality of heavily back loaded noun phrases in written data, using *process* as an example. In the BASE corpus, by contrast, heavy back loading of nouns is not common – we find *decision making process*, *selection process*, *production process* but little else – and students can look at process and other common nouns and investigate pre and post modification of these nouns, and which verbs are used with the nouns, for example. In addition, they can find out which relations are typical of which disciplinary domain.

Further work can be done by looking at frequent lexical bundles and P-frames (that is, fixed or variable word sequences). Space does not allow

for further elaboration here, but it should also be noted that it is not sufficient to recognise only the orthographical form of the word in context, but also to understand the phonological aspects. Students need to know the pronunciation of each item and be able to recognise it in the flow of speech. Field (2011) observes that L2 students up to CEFR C2 have difficulty in decoding what lecturers say; it therefore makes sense to create activities that help students to decode more effectively. One starting point is to familiarise learners with the ways that lecturers speak. Listening activities can be designed in such a way that attention is given to items on the VALL list. For example, learners can be played a 4-5 minute recording of authentic lecture material which is paused at predetermined points of the recording, at which target vocabulary occurs, and at that point learners are asked to write down the last few words that they have heard (this activity is adapted from Field, 2011).

## 8. Conclusions

In this study, I have evaluated four sets of baseword lists for building a word list for listening to academic lectures. The gold standard general word list, the General Service List, was rejected as the basis for forming a word list for academic lecture listening on the grounds that it is outdated and it does not provide a good enough coverage of the most frequent words in the academic lectures.

Three alternatives were then evaluated against the GSL and of these the BNC 1K and 2K lists were judged to be the best choice. A key factor in this decision was that the choice should fit the purpose: in this case, we are concerned with identifying highly frequent vocabulary for use in EAP higher education teaching contexts where students are young adults or older. The BNC word lists provide a rounded coverage of literary, media, academic and spoken language and therefore constitute the best basis for an academic lecture word list, and perhaps also for the construction of academic word lists in general.

The BNC first two thousand words plus the VALL List provided approximately 93% coverage of the tokens in the BASE and MICASE corpus data. Van Zeeland and Schmitt (2013) found that 95% coverage is required for comprehension of spoken discourse. Although their study did not test comprehension of academic lectures, it is possible that the 2200 word families captured, at 93%, and with proper nouns (0.7-0.8%) provide a broad enough lexical knowledge for comprehension, in conjunction with relevant topic specific vocabulary. This, however, needs to be tested empirically.

As explained above, 288 of the AWL families are contained within the BNC 1K and 2K. The VALL 200, in turn, has 105 word families which are also in the AWL. One implication of this is that the AWL does provide a good indication of the lexis of academic lectures (393 word families in the AWL are highly frequent in academic lectures) but a second and important point is that the GSL does not give an accurate indication of what constitutes high frequency vocabulary in modern contexts. If the concept of the first two thousand word families is to have a firmer foundation in reality then it is time to base word lists, particularly in the field of EAP, on the BNC frequency lists rather than the classic but dated GSL. In other words, the AWL has been confirmed as a reliable index of the vocabulary of academic discourse, both written and spoken, but the value of using the GSL as a basis for generating academic word lists is doubtful.

This study has worked with the best resources available but these are still limited. The corpora contain transcripts from 160 lectures in the UK and 62 in the US, which is a very small sample given the immense numbers of lectures taught every day around the world, and obviously the word list can be improved through testing on larger quantities of data.

Since beginning the study, two new general service lists have been published. One has been developed by Charles Browne, Brent Culligan and Joseph Phillips, at Meiji Gakuin University in Japan, using the two billion word Cambridge English Corpus, and working under the guidance of Paul Nation. Their list is available at www.newgeneralservicelist.org/. The second has been created by Vaclav Brezina and Dana Gablasova and is introduced in Brezina and Gablasova (2013). I have not tested these new General Service Lists on the BASE corpus, against the other lists, and there is certainly a possibility that either of the new lists is as good as the BNC 1K and 2K used here. This does not invalidate the argument, however, that the Michael West GSL needs to be retired.

The aims of the study were to develop a robust procedure for building a spoken academic word list, and also to provide information that will assist teachers, learners and materials developers. The resulting word lists are provided as a resource to be drawn on when selecting vocabulary items for study or for pedagogical treatment, and also for assessing the vocabulary load of listening material for teaching and testing purposes. Corpus investigation activities using the BASE corpus in Sketch Engine have been proposed. These activities aim to sensitise learners to the ways that frequent items collocate with other words (particularly with other high frequency words), and the patterns that they occur in. It is important to remember that the list is constructed according to frequency and range

information primarily, and that other criteria for selection should also be considered by teachers and materials writers, such as coverage, availability, opportunism and centres of interest (White 1988, 48-50).

## Notes

[1] The recordings and transcriptions used in this study come from the British Academic Spoken English (BASE) corpus. The corpus was developed at the Universities of Warwick and Reading under the directorship of Hilary Nesi and Paul Thompson. Corpus development was assisted by funding from BALEAP, EURALEX, the British Academy and the Arts and Humanities Research Council.

[2] The Range programme and BNC, GSL, AWL, BNC/COCA word lists were downloaded from https://www.victoria.ac.nz/lals/about/staff/paul-nation. The BNL was downloaded from Bare Naked Lexis Wiki: http://www.editthis.info/thebnl/

[3] Setting the frequency to 8 is an arbitrary decision but it suggests that the type will occur in 8/40, or one in five texts in that disciplinary domain. For the AWL, Coxhead decided that a member of a word family had to occur at least 10 times in each of the four main sections of the corpus (of 3.5 million words) and in 15 or more of the 28 subject areas; the BASE corpus is smaller than the Academic Corpus and thus 8 instances per domain is a reasonable condition.

## References

Adolphs, Svenja. 2006. *Introducing electronic text analysis: A practical guide for language and literary studies*. London: Routledge.

Adolphs, Svenja, and Norbert Schmitt. 2003. Lexical coverage of spoken discourse. *Applied Linguistics* 24: 425-438.

Biber, Douglas. 2006. *University language: A corpus-based study of spoken and written registers*. Amsterdam: John Benjamins.

Billuroğlu, Ali, and Steve Neufeld. 2007. *BNL 2709 The essence of English* (4th ed.). Nicosia: Rüstem Kitabevi.

Brezina, Vaclav, and Dana Gablasova. 2013. Is there a core general vocabulary? Introducing the New General Service List. *Applied Linguistics* 1-23. Open Access, first published online August 26, 2013 doi:10.1093/applin/amt018

Coxhead, Averil. 2000. A new academic word list. *TESOL Quarterly* 34(2): 213-238.

—. 2011. The Academic Word List ten years on: Research and teaching implications. *TESOL Quarterly* 45(2): 355-362.

Eldridge, John. 2008. “No, there isn’t an ‘academic vocabulary,’ but . . .”: A reader responds to K. Hyland and P. Tse’s ‘Is there an ‘academic vocabulary’? *TESOL Quarterly* 42(1): 109-113.

Faucett, Lawrence, Harold Palmer, Edward Thorndike, and Michael West. 1936. *Interim report on vocabulary selection*. London: P.S. King and Son, Ltd.

Field, John. 2011. Into the mind of the academic listener. *Journal of English for Academic Purposes* 10: 102-112.

Gilner, Leah. 2011. A primer on the General Service List. *Reading in a Foreign Language* 23(1): 65-83.

Hancioğlu, Nilgun, Steve Neufeld, and John Eldridge. 2008. Through the looking glass and into the land of lexico-grammar. *English for Specific Purposes* 27(4): 459-479.

Heatley, Alex, Paul Nation and Averil Coxhead. 2002. *Range: A program for the analysis of vocabulary in texts* [software]. Downloadable from http://www.victoria.ac.nz/lals/about/staff/paul-nation

Hunston, Susan. 2002. *Corpora in applied linguistics*. Cambridge: Cambridge University Press.

Hyland, Ken, and Polly Tse. 2007. Is there an "academic vocabulary"? *TESOL Quarterly* 41(2): 235-253.

Laufer, Batia, and Geke Ravenhorst-Kalovski. 2010. Lexical threshold revisited: Lexical coverage, learners' vocabulary size and reading comprehension. *Reading in a Foreign Language* 22(1): 15-30.

Leech, Geoffrey, Paul Rayson, and Andrew Wilson. 2001. *Word frequencies in written and spoken English: Based on the British National Corpus*. Harlow: Longman.

Murphy, Gregory. 2004. *The big book of concepts*. Cambridge, MA: MIT Press.

Nation, Paul. 1990. *Teaching and learning vocabulary*. New York: Newbury House.

—. 2004. A study of the most frequent word families in the British National Corpus. In *Vocabulary in a second language: Selection, acquisition, and testing*, ed. Paul Bogaards and Batia Laufer, 3-13. Amsterdam: John Benjamins.

—. 2006. How large a vocabulary is needed for reading and listening? *Canadian Modern Language Review* 63(1): 59-82.

Nation, Paul, and Robert Waring. 1997. Vocabulary size, text coverage and word lists. In *Vocabulary description, acquisition and pedagogy*, ed. Norbert Schmitt and Michael McCarthy, 6-19. Cambridge: Cambridge University Press.

Nation, Paul, and Stuart Webb. 2011. *Researching and analyzing vocabulary*. Boston: Heinle Cengage Learning.

Nesi, Hilary, and Sheena Gardner. 2012. *Genres across the disciplines: Student writing in higher education*. Cambridge: Cambridge University Press.

Nesi, Hilary, and Paul Thompson. 2006. *The British Academic Spoken English corpus manual*. Available at http://www.coventry.ac.uk/base

Neufeld, Steve, and Ali Billuroğlu. 2006. *The Bare necessities in lexis: A new perspective on vocabulary profiling*. Retrieved September 16, 2014, from http://www.lextutor.ca/vp/bnl/BNL_Rationale.doc

Newton, Jonathan, and Paul Nation. 1997. Vocabulary and teaching. In *Second language vocabulary acquisition*, ed. James Coady and Thomas Huckin, 238-254. Cambridge: Cambridge University Press.

Schmitt, Diane, and Norbert Schmitt. 2005. *Focus on vocabulary: Mastering the academic word list*. London: Pearson Education.

Schmitt, Norbert, Xiangying Jiang, and William Grabe. 2011. The percentage of words known in a text and reading comprehension. *Modern Language Journal* 95(1): 26-45.

Schonell, Fred, Ivor Meddleton, and B. A. Shaw. 1956. *A study of the oral vocabulary of adults*. Brisbane: University of Queensland Press.

Simpson, Rita, Sarah Briggs, Janice Ovens, and John Swales. 2002. *The Michigan corpus of academic spoken English*. Ann Arbor, MI: The Regents of the University of Michigan.

Sinclair, John McH. 1991. *Corpus, concordance, collocation*. Oxford: Oxford University Press.

Sinclair, John McH., and Antoinette Renouf. 1988. A lexical syllabus for language learning. In *Vocabulary and language teaching*, ed. Ronald Carter and Michael McCarthy, 140-160. Harlow: Longman.

Tribble, Christopher. 2000. Genres, keywords, teaching: Towards a pedagogic account of the language of project proposals. In *Rethinking language pedagogy from a corpus perspective: Papers from the third international conference on teaching and language corpora*, ed. Lou Burnard and Tony McEnery, 75-90. Hamburg: Peter Lang.

Van Zeeland, Hilde, and Norbert Schmitt. 2012. Lexical coverage in L1 and L2 listening comprehension: The same or different from reading comprehension? *Applied Linguistics* 34(4): 457-479.

Wells, Linda. 2007. *Vocabulary mastery 1: Using and learning the academic word list*. Ann Arbor, MI: University of Michigan Press.

Wells, Linda and Gladys Valcourt. 2008. *Vocabulary mastery 2: Using and learning the academic word list*. Ann Arbor, MI: University of Michigan Press.

Wells, Linda and Gladys Valcourt. 2010. *Vocabulary mastery 3: Using and learning the academic word list*. Ann Arbor, MI: University of Michigan Press.

West, Michael. 1953. *A general service list of English words*. London: Longman, Green & Co.

White, Ron. 1988. *The ELT curriculum: Design, innovation and management*. Oxford: Blackwell.

Willis, Dave. 1990. *The lexical syllabus*. London: Collins ELT.

Xue, Guoyi, and Paul Nation. 1984. A university word list. *Language Learning and Communication* 3: 215-229.

Zipf, George. 1945. The meaning-frequency relationship of words. *Journal of General Psychology* 33: 251-256.

## Appendix: The 200 word family headwords in the Vocabulary for Academic Lecture Listening word list

abstract
academy
acquire
acute
adapt
agriculture
alpha
approximate
arise
asia
atmosphere
audit
bacterium
barrier
behave
belief
biology
bond
cancer
capitalism
carbon
circulate
classic
coin
colony
column
complement
component
compose
conclude
conduct
consent
consequence
constrain
construct
contrast
convention
core
correspond
crop
crucial
curve
decline
decrease
dense
derive
differ
dimension
distinguish
dominant
elastic
eliminate
email
emerge
empirical
equation
equilibrium
equivalent
error
essay
evaluate
explicit
explore
export
expose
external
failure
false
feedback
fibre
fluid
formula
framework
frequent
fundamental
gender
gene
genuine
global
handout
height
hence
host
hypothesis
illustrate
imply
import
importance
index
infer
inflate
informal
input
integrate
intellectual
intense
interact
interfere
intervene

invent
isolate
laboratory
layer
legislate
linear
literature
loop
manipulate
matrix
mechanism
media
military
mixture
mode
module
molecule
motive
multiple
network
notion
novel
nuclear
objective
organic
organism
origin
outcome
outline
output
overhead
oxygen
paradigm
parallel
perceive
perception
perspective
pest
phase
phenomenon
philosophy
phrase
plot
polar
possess
poverty
predict
presence
primary
prior
professor
protein
quantity
radiate
radical
random
rapid
ratio
rational
regime
replicate
reproduce
review
seminar
sensitive
sequence
so-called
somewhat
species
statistic
stimulate
straightforward
strain
strict
subsequent
substance
substitute
summary
surround
sustain
symbol
task
tense
textbook
theme
transform
transition
translate
transmit
trend
tutor
underlie
universe
upper
vague
versus
vertical
visual
web
welfare
yield

# CONTRIBUTORS

**Laurie Anderson** is Professor of English at the University of Siena (Italy). Her current research is articulated into two complementary strands, both engaging critically with issues related to the use of English as a lingua franca (ELF) in Continental Europe: the investigation of the multilingual practices of internationally-mobile scholars and the use of ELF in interactions involving migrant patients. Recent publications in these areas include: "Publishing strategies of young, highly mobile academics: The question of language in the European context" (*Language Policy*, 2013) and "Code-switching and coordination in interpreter-mediated interaction" (in Baraldi/Gavioli eds., *Coordinating Participation in Dialogue Interpreting*, Benjamins, 2012). She collaborates with the Max Weber Post-doctoral Programme at the European University Institute (Florence) and is a founding member of the FIESOLE Group, a network of applied linguists from various European institutions dedicated to developing a reflexive, transnational approach to training for academic practice in English.

**Silvia Bernardini** is Associate Professor of English Language and Translation at the University of Bologna (Italy), Department of Interpreting and Translation, where she coordinates the Master's in specialised translation. She has taught specialised translation, translation technology and English linguistics courses. Her research interests include corpus-based translation studies (of phraseology in particular), English as a Lingua Franca in institutional settings, construction and use of corpora for professional, pedagogic and research purposes. From 2001 to 2014 she has edited the international journal *Languages in Contrast* (Benjamins).

**Geneviève Bordet** is Associate Lecturer at Paris Diderot University (France), Department of Applied Linguistics. She is co-directing a Master in specialised translation and teaches terminology and information research for specialised translation. Her research focus is set on linguistic analysis of the lexical items participating in the construction of a credible "academic voice" in scientific discourse, as exemplified for instance by PhD abstracts, the analysis of which she conducts in double contrastive perspective, by confronting English and French languages and various

scientific disciplines. This research topic involves studies of the text-structuring role of lexical chains and the use of anaphoric labeling nouns determined by "this". Her recent research work also centres on corpus analysis-based methods for teaching specialised translation.

**Maggie Charles** is Tutor in English for Academic Purposes at Oxford University Language Centre (UK), where she teaches academic writing to graduate students. Her research interests include the pedagogical applications of corpus linguistics, the study of stance/evaluation and discipline-specific discourse and she has published widely in these areas, including *Academic Writing: At the Interface of Corpus and Discourse* (Continuum, 2009), co-edited with Diane Pecorari and Susan Hunston and *Corpora, Grammar and Discourse* (Benjamins, forthcoming), co-edited with Nicholas Groom and Suganthi John. Recently she was the consultant on academic writing for *Oxford Advanced Learner's Dictionary* (2010) and *Oxford Learner's Dictionary of Academic English* (2014).

**Giuliana Diani** is a Lecturer in English Language and Translation at the University of Modena and Reggio Emilia (Italy). She holds an MA in Language Studies from the University of Lancaster (UK) and a PhD in English Linguistics from the University of Pisa (Italy). She has worked on various aspects of discourse analysis and EAP, with special reference to metadiscourse and evaluative language. Her recent work centres on language variation across academic genres, disciplines and cultures through the analysis of small specialised corpora. Her recent publications include: *Variation and Change in Spoken and Written Discourse: Perspectives from Corpus Linguistics* (co-edited with Julia Bamford and Silvia Cavalieri, Benjamins, 2013); *Reviewing Academic Research in the Disciplines: Insights into the Book Review Article in English* (Officina Edizioni, 2012); *Academic Evaluation: Review Genres in University Settings* (co-edited with Ken Hyland, Palgrave, 2009).

**Adriano Ferraresi** is a Postdoctoral Fellow at the Department of Interpreting and Translation (University of Bologna at Forlì, Italy), where he also teaches courses in English->Italian translation and translation technology. He holds a Doctorate in English for Special Purposes from the University of Naples "Federico II" and his research interests are in the areas of phraseology, especially from a corpus-based perspective, institutional academic English and development of reference and specialised corpora from the web.

**Maria Freddi** is Associate Professor of English Language and Linguistics at the University of Pavia (Italy), where she currently teaches courses in English grammar and text, corpus linguistics and English for Academic Purposes. As of this academic year she will also be teaching technical English to graduate students of Engineering. Her main areas of writing and research are corpus linguistics and quantitative methods in language description, the rhetoric of science, including English for Specific and Academic Purposes, and the role of descriptive grammars in the teaching and learning of English as a second or foreign language. She is interested in the pedagogy of English directed at students who specialise in Modern Languages and Linguistics as well as non-specialist language learners who need English as a tool for professional communication and knowledge dissemination in international contexts.

**Christopher Gledhill** is Professor of English Linguistics at Paris Diderot University (France). He has taught and published on discourse analysis, interlinguistics, phraseology, systemic functional grammar and specialised translation. Previously, he was a lecturer in French Linguistics at the universities of Aston and Saint Andrews. He is currently working on a corpus-linguistic analysis of English as a Lingua Franca, with special reference to phraseology and collocational patterns.

**Šarolta Godnič Vičič** is a Senior Lecturer in English at the Faculty of Tourism Studies - Turistica, University of Primorska (Slovenia). Her main research and teaching interests include EAP and ESP, discourse analysis as well as genre variation and change. She is particularly interested in variation and change in research articles, academic publishing, tourism discourse, corpus linguistics and foreign language acquisition. She has extensive experience in developing teaching materials for ESP undergraduate study programmes. Her publications comprise articles and chapters on the above topics and coursebooks for English for tourism purposes. She currently serves as an Associate Editor of the LSP journal *Scripta Manent*.

**Mojca Jarc** works as Lector in English and in French Language at the Faculty of Social Sciences, University of Ljubljana (Slovenia). She is Editor of *Scripta Manent*: *Journal of the Slovene Association of LSP Teachers* and a member of the board of reviewers of *Recherche et Pratiques Pédagogiques en Langues de Spécialité: Cahiers de l'APLIUT*. Her research interests lie in the field of applied linguistics. Her research centres on various aspects of LSP teaching, learning, and second-language

writing. She has published articles and chapters on problem based learning in ESP, terminology of social sciences and publication practices of Slovene social sciences authors. She has also co-authored two LSP coursebooks for social sciences students.

**Rosa Lorés-Sanz** is a Senior Lecturer in the Department of English and German Studies at the University of Zaragoza (Spain), where she teaches Linguistics and Translation. She holds an MPhil in Translation from the University of Salford (UK) and a PhD in English Linguistics from the University of Zaragoza (Spain). She has edited several books and has published articles in national and international journals on pragmatics and translation, and corpus and contrastive studies (English-Spanish) applied to academic and specialised discourses. Her present research focuses on the exploration of rhetorical and lexicogrammatical features in written academic genres (abstracts, research articles and book reviews) mainly from a cross-cultural and cross-linguistic perspective, as well as on the use of English as a lingua franca by Spanish academics. She is a member of the research group InterLAE (www.interlae.com).

**Stefania M. Maci** is Professor of English Language and Translation at the University of Bergamo (Italy). She is a member of CERLIS, CLAVIER, BAAL, and AIA. Her research is focused on the study of the English language in academic and professional contexts. Her most recent publications include: the monographs *Tourism Discourse: professional, promotional, digital voices* (2013); the co-edited volume *Genre Variation in Academic Communication. Emerging Disciplinary Trends* (2012); and the papers: "Institutional popularization of medical knowledge: The case of pandemic influenza A (H1N1)"(2014); "The popularisation of scientific discourse for the Academia from a diachronic perspective: The case of Nobel lectures" (2014); "Tourism as a specialised discourse: The case of normative guidelines in the European Union" (2012); "Arbitration in action: The display of arbitrators' neutrality in witness hearings" (2012); "The Discussion section of medical research srticles: A cross cultural perspective" (2012); "The genre of medical conference posters" (2012).

**Giuseppe Palumbo** is a Lecturer of English Language and Translation in the Department of Legal, Language, Interpreting and Translation Studies at the University of Trieste (Italy), where he is also the Director of the Language Centre. He holds a PhD from the University of Surrey and specialises in technical and scientific translation. His research interests include translation technology, corpus linguistics, English for academic

purposes and the use of English in international institutional settings. He has published on terminology, the design of translator training curricula and institutional translation, and is also the author of *Key Terms in Translation Studies* (Continuum, 2009).

**Michele Sala**, PhD (University of Bergamo), MA (Youngstown State University, Ohio), is a Researcher in English Language and Translation at the University of Bergamo (Italy), Department of Foreign Languages, Literatures and Communication. His research activity and major publications deal with language for specific purposes and, more specifically, the application of genre and discourse analytical methods to a corpus-based study of legal-academic discourse and the analysis of the linguistic, textual and pragmatic aspects of legal translation. He has also published in the field of academic discourse (*Persuasion and Politeness in Academic Discourse*, 2008), pragmatics and cognitive linguistics (*Differently Amusing* 2012).

**Paul Thompson** is the Director of the Centre for Corpus Research at the University of Birmingham (UK). His research interests are in academic and other specialised discourses, in the linguistic aspects of human-computer interaction, in uses of educational technologies in language learning, and in the exploitation of corpus resources and methodologies in learning about language. He is currently Principal Investigator on an ESRC-funded project 'Investigating Interdisciplinary Research Discourse: the case of Global Environmental Change', a research collaboration with the scientific publisher, Elsevier. He is a former Secretary of the British Association for Applied Linguistics (BAAL), and has been Co-Editor of the *Journal of English for Academic Purposes* since 2009. He has published many papers and edited several collections on corpus and applied linguistics. Most recently he co-edited, with Ana Diaz-Negrillo and Nicolas Ballier, "Automatic Treatment and Analysis of Learner Corpus Data" (Benjamins, 2013).

# INDEX


abstracts 19-21, 30, 46-47, 49, 51, 54, 58, 61-63, 103, 105-106, 120, 122-123, 127, 129, 133-134, 137, 142, 181-182, 268, 276
academic writing 130, 151, 153, 177- 178, 186-187, 268, 275-279
  science writing 11, 18, 26, 30, 35-36
Ädel, A. 2, 151, 185
altercasting 201-202, 214, 217-218
argumentation 62, 105-106, 120, 136, 143, 145
Bernardini, S. 229-230, 247, 267, 286
Bhatia, V. K. 1, 49, 104, 118, 182, 187, 247
Biber, D. 2, 13, 15, 43-45, 85, 92, 105, 152, 154, 227, 253, 317, 319, 323
Bondi, M. 2, 44-46, 49, 52, 79, 104-107, 151-153, 157, 161, 163-165, 247, 266, 285-286, 301
book review article 1, 3-4, 151-152, 158, 162-164, 247
British National Corpus (BNC) 16-17, 23-24, 84, 135-137, 233, 319, 324-326
British Academic Spoken English (BASE) 322, 337
Bunton, D. 1, 267, 269-273, 276-277
Charles, M. 2, 44, 79, 152, 267, 275, 278, 280, 286-287, 295
collocation 3, 14, 44-45, 84, 106, 227, 232-238
  collocational cascade 20-22
Connor, U. 173-177, 225
corpus methods 265, 286
  concordances 20, 154, 297,
  keywords 16, 43-49, 52-63, 84-85, 98, 105, 154, 276
  wordlists 53, 154, 286, 317
  academic wordlist 317
course design 285, 302
Coxhead, A. 286, 318-323, 330
Dahl, T. 79, 151-152
discipline 45-47, 79-80, 103-105, 151, 179-181, 184, 202, 227, 269, 287-290, 295, 318
  humanities 179, 287-290, 322, 334
  economics 105, 152-153, 155-164, 181, 199
  business 152-164, 181-185
  sociology 79-88, 96-98, 179
DIY corpora 275
English for Academic Purposes (EAP) 1, 152, 173-174, 187, 198, 219, 239, 279, 285, 292, 296, 317
  EAP reading 287, 302
  EAP teaching 186, 276, 285, 301, 318, 331
  EAP writing 176, 268, 277, 286
  EAP pedagogy 2, 265, 279, 288
English as a lingua franca (ELF) 3, 5, 197, 220, 228, 247
epistemic 132, 186, 313
evidentials 182-183
Flowerdew, J. 187, 276, 286, 294
Flowerdew, L. 2, 265-267, 276, 286, 289
General Service List 6, 318, 320, 335-336
genre analysis 1-6, 175-177, 269

Gledhill, C. 11, 15-16, 19-20, 43-44, 79, 84, 97
Gotti, M. 2, 105, 118, 248
Groom, N. 13, 22, 84-85, 152
Halliday, M. A. K. 25, 44, 58, 61
Hoey, M. 22, 43-44, 83, 267, 276
Hunston, S. 2, 19, 89-90, 97, 147, 278, 332
Hyland, K. 1-2, 13, 44, 46-47, 49, 63, 79, 82, 103-106, 118, 151-152, 154, 163, 181-182, 187, 225, 285-286, 289-290, 293, 321, 334
identity 162, 177, 186, 188, 201, 203, 206, 209, 211, 220
if-conditionals 131, 136
intercultural 3-4, 173-177, 181, 183-187, 200, 203, 214-215, 249
institutional academic English 225-226, 229, 232
  institutional communication 245, 247-248, 250, 258-259
Jenkins, J. 197-198, 228
lecture 6, 317-319, 322-335
lexico-grammar 274, 293, 252, 265, 275, 303
  lexico-grammatical patterns 3, 13-16, 19-20, 28-29, 31, 36, 43, 186, 304
  extended lexico-grammatical patterns 22
  grammatical items 3, 11-16, 18-22, 36, 332
  lexical chains 52, 55, 61, 63
  lexical association measures 232, 235, 237
Lorés Sanz, R. 1, 79-80, 104, 182, 185
Mauranen, A. 151, 154, 174-175, 178, 186, 189, 197, 219, 228, 240, 247-250
Membership Categorization 199-200
metadiscourse 151, 181-182, 185-186
Mur Dueñas, P. 79, 181-184, 188
Nation, P. 291, 317-321, 323, 325, 332
needs analysis 288-289
Nesi, H. 152, 285-286, 331, 337
nominalisation 295-296, 300
  noun phrases 254, 258-259, 296-297, 334
Pérez-Llantada, C. 178-179, 181, 185-187, 190
phraseology 2-5, 11, 19-20, 22, 24, 32, 79, 105, 151-154, 186, 225-226, 232, 240-241, 292, 295
reflexivity 151, 154
Renouf, A. 12-13, 22, 43-44, 331
research articles 3-4, 12, 16, 19, 79, 132, 136-137, 153, 158-165, 179, 188, 275
  journalistic article 23
research presentations 197-199
rhetorical sections 16, 18, 186
Scott, M. 13, 16, 84-85, 97, 106, 134, 154
Sinclair, J. McH. 13, 15, 20, 22, 43-44, 154, 233, 331, 334
protasis 132-138, 141-145
self-categorization 202-207, 211
semantic associations 86, 88, 91, 96-98
Swales, J. 1, 43, 46, 49, 51, 63, 80, 83, 103-104, 118, 122, 130, 137, 163, 187, 225, 228, 247, 265-266, 269, 275-276, 286-287, 294-296, 300, 302
syllabus design 287, 302
thematization 5, 115, 118, 120
thesis 267-273, 276
Thompson, P. 1, 273, 286, 337
translation 246-250
university websites 5, 226, 228-230, 245-247, 250-251, 258
university internationalization 226, 187, 198, 285, 291
variation 2-3, 18, 27, 36, 37, 43-47, 61, 79-80, 82, 96, 105,116, 118,

151, 254, 269-270, 290, 294, 297, 300, 328
(cross)disciplinary variation 63, 98, 153-154, 157, 266
intradisciplinary variation 3, 80, 83-86, 97-98
lexical variation 45-49, 55, 57, 60, 62-63
(cross)cultural variation 79, 175
vocabulary 1-2, 248, 289, 291-295, 300-303, 317-321, 323, 325-326, 330-332, 335-336
web-as-corpus 226, 229-230
word families 303, 318-321, 323-331, 335-336